92810

Technical Assessment of Open Data Platforms for National Statistical Organisations

18 October, 2014

World Bank Group

1818 H Street NW
Washington DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

For questions or comments concerning this working paper, please contact Timothy Herzog (therzog1@worldbank.org) or Amparo Ballivian (aballivian@worldbank.org).

Contents

1 Executive Summary

National Statistics Offices (NSOs) have the potential to play a pivotal role in the implementation of open data initiatives. As producers and curators of data, the objective of making high quality data more accessible and usable is consistent with their guiding principles.

NSOs indicate, in research conducted in support of this report, that one of the difficulties they encounter is that the technology they use to publish - or electronically distribute - data for public use is not compatible with open formats. They also indicate that common software packages used for open data portals do not accommodate the data formats and metadata they produce.

This research report is intended to provide a better understanding and assessment of the technical issues related to data dissemination tools that NSOs use (or could use) to distribute data to the public under an open data initiative. The report defines a list of key criteria and evaluates relevant technology products according to those criteria.

Two key concerns related to data dissemination products are addressed:

1. Can such products designed primarily for NSOs satisfy requirements for an open data initiative?
2. Can such products designed primarily for open data satisfy the requirements of NSOs?

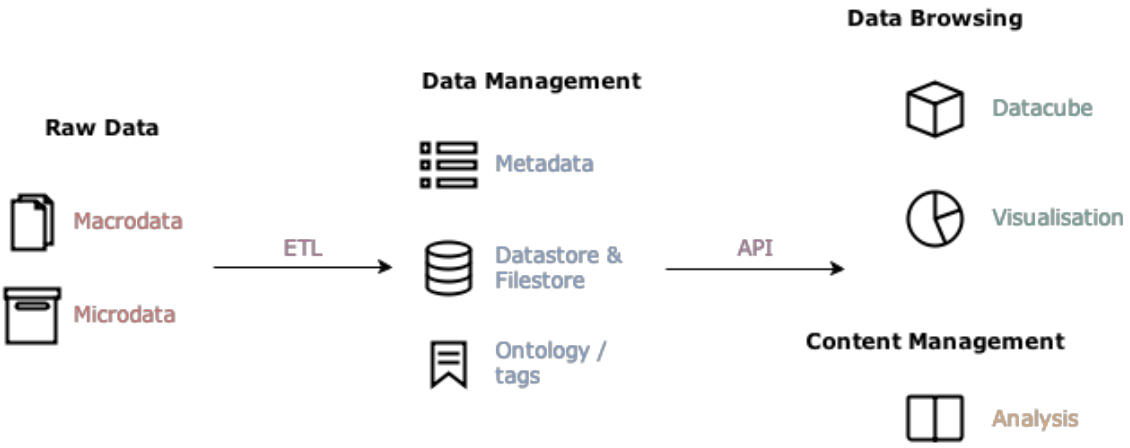
The main audiences for this report are NSO staff, particularly managers and directors, seeking to better understand technology issues relevant to dissemination of open data. Software developers, donor agencies and consultants wishing to strengthen open data systems and producers will also find it useful.

This report is limited to a technical discussion of data dissemination platforms; it does not cover up-stream data production and curation issues. The report also does not cover a host of non-technology issues that are nonetheless essential for NSOs to address, such as user engagement, privacy protections, resource constraints and data quality. Many of these issues are explored in depth in *Open Data Challenges and Opportunities for National Statistical Offices*[1].

The term “open data” is generally understood to be data that are made available to the public free of charge, without registration or restrictive licenses, for any purpose whatsoever (including commercial purposes), in electronic, machine-readable formats that ensure data are easy to find, download and use.

Data reuse, both by data experts and the public at large, is key to creating new opportunities and benefits from government data.

A visual representation of the various data and software components which form part of an overall open data platform are presented as follows:



These components are often served by different software systems. Deployments should ensure system interoperability. Different software systems shouldn't result in duplicated functionality or needless requirements for manual data conversion for interoperability where such systems share common data resources.

Open data distribution software systems were not originally designed to serve NSOs, nor were NSO systems originally designed for open data. While some data publication platforms offer Extract, Transform and Load, Business Intelligence and Content Management functionality, this report concentrates on Data Discovery and Publication as the core requirements for any open data distribution software.

Besides the specific components of an open data system, NSOs must also consider the metadata standards they need to support, as well as their procedures for disseminating microdata. Dissemination of microdata is of particular concern to NSOs because of the need to protect privacy of respondents and other concerns. It is common for NSOs to have a separate set of dissemination policies - and separate dissemination platforms - for microdata, and there are already platforms and approaches that are well suited for this purpose.

Proprietary file formats - SAS, STATA, SPSS and so forth - must also be available in more generally accessible machine-readable formats to meet open data best practices. Data dissemination in proprietary formats does not preclude dissemination in open formats and vice versa.

Below is a complete list of assessment criteria for the software platforms described in this report in the order in which they are presented:

Descriptive metadata:	Corresponds to external metadata and typically used for discovery and identification, as information used to search and locate an object such as title, author, subjects, keywords, publisher.
Machine-readable:	Data available as machine-readable structured data in a non-proprietary format can be easily read and used by software systems without human interpretation.
Anonymous access:	Users can search for and access data and metadata without having to identify themselves, create a user account, or receive advance permission.

Data licences:	Data licenses (terms of use) associated with each dataset are clearly presented to the user and permit reuse and republication of that data in any alternative form.
Data attribution:	Users can cite, attribute, and link to datasets, and contact data owners if they have questions.
Search:	Search results should return focused summaries on datasets, along with keywords which aid classification, and the option of reviewing the data online to assess its content.
Application Programming Interface (API):	Platforms make their contents available to external systems by supporting programmatic queries and access to metadata and resources.
Uniform Resource Identifiers (URI):	Platforms make datasets available at persistent URIs that never change, allowing them to be externally referenced reliably.
Harvesting:	An automated and autonomous mechanism for ETL of known data from known web addressable locations into a single database or datastore.
Federation:	A meta-database management system, which transparently maps multiple autonomous database systems into a single federated database, allowing discrete data publishing systems to be integrated, yet operate independently.
Public Documentation:	Data platforms provide comprehensive information for developers and the general public on how their platform works. Such documentation should be updated with each new software release.
Standards-based:	Platforms are consistent with emerging standards recognised by the W3C especially as regards metadata, RDF, and hypercubes.
Structural metadata:	Corresponds to internal metadata about the structure of database objects such as tables, columns, keys and indexes.
Online Analytical Processing (OLAP) hypercube or cube:	Supports the ability to analyse multidimensional data interactively from multiple perspectives, including the ability to consolidate, drill-down, or slice and dice data.
Data Endpoints:	Provides structured data endpoints return data in predictable ways. These can be as simple as a known type of serialisation format while more complex implementations permit the data to be queried, filtering or refining the dataset prior to download.
Visualisation:	Provides tools to present data as common charts, maps or perform more complex statistical analysis.
UX & S/W Extensibility:	Permits sufficient template and layout customisation to provide a consistent user-experience and provide a common look and feel across all NSO online services.

During the research stage for this report, custom software implementations were also assessed along with commonly-used commercial software. Some organizations, for instance the US Census Bureau, have developed their own systems. In large part, such work was initiated long before there was an open data movement, let alone software to support it. Such custom software can often become more generally accepted amongst closely-related NSOs and be released more formally. This is true in the case of Nesstar and PX-Web.

A representative sample of the most commonly used open- and statistical data publication software platforms was evaluated relative to these components:

Software platform	Open Data-Specific Criteria												NSO-Specific Criteria				
	Descriptive Metadata	Machine-readable	Anonymous access	Data licences	Data attribution	Search	Open API	Static URI	Harvesting	Federating	Public Documentation	Standards-based	Structural Metadata	OLAP Hypercubes	Data Endpoints	Visualisation	UX & S/W Extensibility
CKAN	●	●	●	●	●	●	●	●	•	•	●	●			•	•	●
DevInfo	●	●	●	●	●	●	●	●			●		●	●	•	●	•
DKAN	●	●	●	●	●	●	●	●	•	•	●	●			•	•	●
Junar	●	●	●	●	●	●	●	●	•	•	●	●			•	●	●
NADA	●		•	●	●	●		●			•	●	●				●
Nesstar	●	●	●	•	•	●		●			•			●		●	•
OpenDataSoft	●	●	●	●	●	●	●	●	•		●	●			●	●	●
PC-Axis and PX-Web	●	●	●	•	•	•	●	●	•	•	•	•	●	●	●	●	●
Prognoz	●	●	●	•	•	●		●						●		●	•
Semantic MediaWiki	●	●	●	●	●	●	●	●			●	●			•		●
Socrata	●	●	●	●	●	●	●	●	•	•	●	●			●	●	●
Swirrl	●	●	●	●	●		●	●	•	•	●	●			•		●

Where ● offers a complete solution and • offers a partial, or incomplete, solution.

Open data software has led to a number of different approaches to resolving requirements for data publication, engagement and reuse. Many of these do not promote interoperability between platforms and new problems are being raised.

This report makes the following recommendations to improve the overall utility of data publication platforms to NSOs and the open data community:

- **Improve technical documentation:** Too little of the available documentation for developing custom components or using the software APIs offers much support to developers. It is essential that these APIs be sufficiently well-documented that they can be of use.
- **Ensure public APIs and endpoints are interoperable:** If vendors are able to agree on a common API – or are able to connect to a common standard – then harvesting from different systems as well as developing applications that integrate a number of unrelated platforms all become extremely straightforward.
- **Presentation of metadata and URIs must conform to W3C standards:** Many software platforms fail to present descriptive metadata which aids discovery and reuse. It needs to be clear what the licensing and reuse policies are, what the data are about, and who is responsible for it. Similarly, data discovery is time-consuming and discrete URIs for each step of the process permits sharing and saving of these states.

- **Natural language search and metadata faceting should be standard:** Free text search along with metadata faceting speeds up data discovery and improves system performance.
- **Structural metadata and hypercube support are core NSO** requirements**:** Individual data tables become more useful when they can be aggregated together and sliced into numerous views for analysis. Offering support for both descriptive and structural metadata (such as adopting the DDI metadata standard), as well as hypercubes (such as the W3C's RDF Data Cube Vocabulary) will improve interoperability and wider utility.
- **Dashboards and visualisations are necessary for user engagement:** User-engagement in the face of vast and complex data is more likely where that data are presented in a tactile and engaging way.
- **Develop data engagement tools for improving data-quality and reuse:** Approaches which promote data quality and reuse include using the data published in visualisations and analysis, showcasing applications developed by users, registering and tracking issues users experience with data quality, and offering users a mechanism to make data requests.

The software platforms considered in this report are not a comprehensive list of those available to NSOs and open data publishers. Even so, many of them come close to meeting all the criteria for both requirements.

Both software developers and NSOs can take heart from this and work together towards building the remaining components. Completing and meeting the recommendations in this report will ensure system interoperability, data migration, and community engagement for data reuse.

2 Introduction to the Technical Assessment

Open data initiatives ensure that public data are freely available in open, electronic, and reusable formats. National Statistical Offices (NSOs) are responsible for maintaining and disseminating a country's official statistics, many of which typically comprise the foundation of any open data program.

NSOs have the potential to play a pivotal role in the implementation of open data initiatives, since the objective of making high quality data more accessible and usable is consistent with their guiding principles.

NSOs may also have existing relationships with other agencies that provide raw data to the statistical system, endowing them with an important role in the local data community. NSOs have expertise in dealing with the many technical issues attendant in publishing public datasets, making them valuable knowledge resources.

Despite these advantages, NSOs do not always feature prominently in government-sponsored open data initiatives.

NSOs indicate that one of the difficulties they encounter is that the technology they use to publish - or electronically distribute - data for public use is not compatible with open formats. They also indicate that common software packages used for open data portals do not accommodate the types of data and metadata they produce.

Myriad technical solutions are available to enable NSOs and other organisations to publish data. From the perspective of NSOs, these products generally fall into one of two groups:

1. Platforms designed specifically for use by NSOs to satisfy the requirements of NSOs and their traditional users (who are typically data professionals with specialised technical experience);
2. Platforms designed to help organisations - particularly government ministries - publish their data under a government-sponsored open data initiative.

These two categories are not necessarily mutually exclusive. However, many current technology solutions were designed with only one of these requirements in mind and the interaction between the two is unclear. NSOs thus face a challenge in developing a strategy and identifying technology solutions to allow them to disseminate their data under an open data initiative.

2.1 Objectives of this report

This research report is intended to provide a better understanding and assessment of the technical issues related to data dissemination tools that NSOs use (or could use) to distribute data to the public under an open data initiative.

This report begins by defining a list of key criteria and the most relevant technology products to assess. The assessment seeks to address two data dissemination product concerns:

1. Can such products designed primarily for NSOs satisfy requirements for an open data initiative?
2. Can such products designed primarily for open data satisfy the requirements of NSOs?

Specific recommendations for features are also identified which would contribute to a common use-case.

This report presents research on a selection of commonly-used open data and NSO publishing software platforms to ascertain and fully understand the design characteristics, core functionalities and business models of each product, and the implications for use by NSOs in disseminating data.

A product matrix, organised as a set of individual assessments, presents a detailed description of each product's features, along with a list of current use cases. We have also investigated research cases where NSOs developed a custom or proprietary platform and the reasons for doing so (i.e., why commercial or open source platforms were deemed insufficient or less desirable). Research included interviews with users and vendors to obtain product information and demonstrations, where possible.

2.2 Who should read this report

- NSO staff, particularly managers and directors, seeking to better understand technology issues relevant to dissemination of open data;
- Developers seeking to better understand the needs of NSOs and other government agencies;
- Donor agencies that support statistical strengthening and capacity building in developing countries;
- Consultants wishing to support NSOs in developing and deploying integrated statistical and open data publication platforms;

2.3 What this report does not cover

There are a wide range of issues and components that form part of an integrated open data initiative. Only those which are technology-based and core to the success and implementation of open data software are covered in this report.

Up-stream data lifecycle issues, such as data production, curation, management, or any other actions which precede data dissemination, are not included. While not necessarily true in practice, this report assumes that NSO data management systems and its data dissemination platform are discrete systems. Discussion of the inter-operability between the publishing platform and the public data discovery systems is an important criterion for NSOs and is included.

This report does not cover a host of non-technology issues that are nonetheless essential for NSOs to address. Many of these issues are explored in depth in Open Data Challenges and Opportunities for National Statistical Offices[2].

Several products have specific features which may be particularly useful in certain contexts, but are not broadly implemented. A comparative analysis of such features is out of scope, including:

- Version control: iterative versions of datasets available for comparative purposes;
- Data collections: organising multiple data resources together as a single set;
- Social media: integration with popular social media services is not unique to data publishing software. Furthermore, open data initiatives require much more comprehensive user-engagement practices than social media implementations typically provide;
- Organisation sub-sites;

This report specifically does not make product recommendations or offer a ranking of the software discussed. It does make suggestions for specific improvements but is not a comprehensive review of all the available software.

3 Overview of data publication and management

"The first rule of data storage: don't store the same data in two different places: you will have problems keeping it consistent," Tim Berners-Lee

Data publication aimed at the public, rather than internal work-streams and data life-cycle management, may seem remote from the day-to-day requirements of an NSO. However, if these public systems are to encourage publication and data reuse then they must integrate well with in-house data process management systems.

There are three generalised systems used to manage online and public versions of research reports, content and research data:

1. **Content Management Systems (CMS):** permit publishing, editing and modifying of qualitative content, as well as providing mechanisms to manage workflows and individual users in a collaborative environment;
2. **Data Discovery Systems (DDS):** are similar to CMS but provide mechanisms to manage the semi-structured quantitative and qualitative data in documents and spreadsheets and offer methods for data publication, discovery and reuse;
3. **Business Intelligence Systems (BIS):** provide a platform for engaging with structured quantitative data to produce custom slices of that data, charts, tables and geospatial representations;

While it is certainly possible for a single, integrated, software system to serve all of these requirements, this is not a common use-case. Data managers usually operate a number of different systems which often require manual data restructuring for availability in each system.

The following sections provide context for the characteristics of systems which support the fulfilment of "data dissemination, user engagement and transparency within an open data initiative, as opposed to up-stream management of data production."[3] Common terms used in the industry are only briefly defined, with references for those requiring a deeper understanding.

3.1 Open Data

The term "open data" is generally understood to be data that are made available to the public free of charge, without registration or restrictive licenses, for any purpose whatsoever (including commercial purposes), in electronic, machine-readable formats that ensure data are easy to find, download and use.

Open data initiatives by public institutions, such as governments and intergovernmental organisations, recognise that such data is produced with public funds and so, with few exceptions, should be treated as public goods.

Data reuse, both by data experts and the public at large, is key to creating new opportunities and benefits from government data. Open data reuse requires two basic criteria:

- **Data must be legally open**, meaning that it is placed in the public domain or under liberal terms of use with minimal restrictions. This ensures that government policies do not create barriers or ambiguities concerning how the data may be used.
- **Data must be technically open**, meaning that it is published in electronic formats that are machine-readable and non-proprietary. This ensures that ordinary citizens can access and use the data with little or no cost using common software tools.

Open data is of particular interest to NSOs for several reasons. NSOs typically manage many of the data products considered high value in government-wide open data initiatives. As governments increasingly develop open data policies, these products (and hence NSOs themselves) may receive greater prominence. The underlying principles of open data are clearly linked to one of the NSOs' fundamental purposes: to make relevant statistics available in ways that are easy for users to access and use. Open data can also create opportunities to increase efficiency of dissemination, improve data quality, lead to the modernisation of administrative records, and raise the public profile of the NSO.

Many NSO products are relatively straight-forward to release as open data, including:

- Statistical products that are already publicly available without restriction; perhaps through printed publications, the NSO's website, or upon request;
- Other vital census and economic statistics at the national and sub-national levels;
- Price and trade data;
- Registers used for drawing statistical samples, for example lists of businesses;
- Official maps of political boundaries, voting districts, infrastructure, and the location of public facilities (schools, government offices, police stations, libraries, etc.);
- Classification systems such as for household consumption or types of industry;

Open data presents opportunities and challenges for NSOs.

3.2 Metadata

The creator of the data would best know what the document is about and should assign keywords as descriptors. This data about the data is called **metadata**. The term is ambiguous, as it is used for two fundamentally different concepts:

- **Structural metadata** correspond to internal metadata (i.e. metadata about the structure of database objects such as tables, columns, keys and indexes);
- **Descriptive metadata** correspond to external metadata. (i.e. metadata typically used for discovery and identification, as information used to search and locate an object such as title, author, subjects, keywords, publisher);

Descriptive metadata permits discovery of the object. Structural metadata permits the data to be applied, interpreted, analysed, restructured, and linked to other, similar, datasets.

Metadata can permit interoperability between different systems. An agreed-upon structure for querying the 'aboutness' of a data series can permit unrelated software systems to find and use remote data.

Beyond metadata, there are also mechanisms for the structuring of relationships between hierarchies of keywords. These are known as **ontologies** and, along with metadata, can be used to accurately define and permit discovery of data.

Adding metadata to existing data resources can be a labour-intensive and expensive process. This may become a barrier to implementing a comprehensive knowledge management system.

Where there are millions of users conducting a high volume of data interactions (in the millions per day), algorithmic systems can assign commonly-used search terms as metadata to data through drawing conclusions from their behaviour (e.g. a particular search result always results in a particular data choice). In the low-frequency user activity of specialist research data, such metadata and ontologies are often developed in advance and assigned manually.

3.3 Microdata and generalised data

Data are often presented as a ubiquitous mass. Statistical data encompasses a range of forms, from questionnaires and individual – personally-identifying – responses, through to aggregated tables of numerical values, analysis and text-driven reports.

In this report we differentiate between:

1. **Microdata:** information at the level of individual respondents, households and businesses, typically through surveys; for example, a national census may collect age, address, education, employment status, etc. from individuals;
2. **Generalised data:** aggregations derived from microdata; for example, the total number of people of a particular education category in the general population;

Dissemination of microdata is an issue of particular concern to NSOs. There is almost always the need, and sometimes, a legal obligation, to protect confidentiality and privacy for the providers of microdata. In some cases, provenance of microdata may lie with external partners (such as academic institutions) with their own policies or restrictions on dissemination. Accordingly, decisions concerning how to manage and disseminate microdata, either as open data or under other policies, is typically made on a case-by-case basis according to the policies and professional judgment of the NSO. However, where provenance concerns do not exist and where privacy issues have been addressed (for instance, through anonymisation techniques), there is not necessarily any reason why microdata should not be released as open data.

NSOs have used a variety of approaches to making microdata publicly available[2]. Some are quite consistent with open data principles; however, many policies place restrictions on who may access the data and how the data may be used.

It is common for NSOs to have a separate set of dissemination policies - and separate dissemination platforms - for microdata that are particularly sensitive than for their other data products. This dual approach may be beneficial for users, since it highlights that certain microdata are provided with additional conditions with respect to data access.

There are already platforms and approaches that are well suited for distributing microdata online.

For example, the **Integrated Public Use Microdata Series (IPUMS)**[4] requires researchers to implement security measures, avoid redistribution of microdata, use microdata only for non-commercial research/education purposes, and not make any attempt to identify the individuals recorded.

The **International Household Survey Network (IHSN)** has developed tools and guidelines to help interested statistical agencies improve their microdata management practices, including a **Microdata Cataloging Tool (NADA)**[5] which is assessed in this report. NADA allows administrators to specify an access policy for each dataset. Policies can range from “Open access” (similar to “open data”) to “Data not available” (metadata only) for each microdata file.

3.4 Proprietary file formats

One of the criteria for open data is the use of machine-readable, non-proprietary electronic data files for data distribution. These formats reduce technical barriers to data access to an absolute minimum for broad categories of users. NSOs commonly distribute data in a variety of formats. Some are considered “open” (such as CSV, XML, text and others) and some are proprietary formats used in data analysis software products (SAS, STATA, SPSS, etc.).

The latter are legitimate even in an open data initiative since these are the software systems used by many professional data users. However, since these formats are not interoperable, the potential for data re-use is limited unless open formats are also supported. Data dissemination in proprietary formats does not preclude dissemination in open formats and vice versa.

If an NSO already distributes data in a proprietary format, it should also distribute in one or more open formats.

3.5 Data structure and linked data

Data is usually stored in a relational database. The approach to such systems is beyond the scope of this report, however, a brief overview is necessary to understand an increasingly popular approach to managing open data on the web.

Structural metadata permits data contained in relational databases to be aligned and joined. This, however, can be slow and inefficient for large and complex datasets.

The World Wide Web Consortium (W3C) develops web standards. Their approach for data interchange on the web is known as the **Resource Description Framework (RDF)**[6]. It has come to be used as a general method for conceptual description or modelling of information that is implemented in web resources, using a variety of syntax formats.

RDF breaks away from the standard relational database and can be thought of as a graph of entity-relationships of the form: subject, predicate and object. The subject (e.g. John) is linked to the object (e.g. Carol) by a predicate (e.g. 'is a friend') and gives rise to the terms **Triple Store** (the three entities being the 'triple') or **Linked Data**.

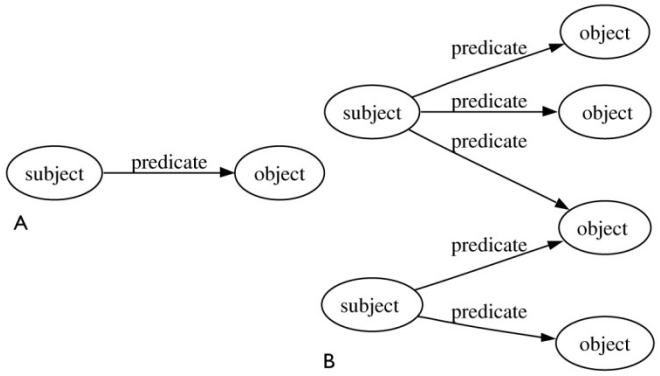


Figure 1: RDF subject, predicate, object model (A) combining to form an RDF graph (B)[7]

The subject and object are also known as **nodes**, while the predicate is an **edge**. A network of nodes linked by edges is called a **graph**.

Numerous implementations of this have resulted in interoperable structured, and machine-readable, metadata systems. There are also, however, numerous legacy approaches to categorising data which have arisen in individual research institutions across the world.

3.6 Software development and deployment

There is tension between the needs of NSOs for standards-compliant but customisable systems, versus proprietary implementations and limited customisability preferred by many vendors.

NSOs have a government mandate to maintain national statistics indefinitely. They legitimately fear that proprietary systems expose them to future data migration costs as such companies go

out of business or are acquired. Numerous custom systems built by NSOs were developed when their existing vendors went insolvent or discontinued software support.

Vendors are concerned that they will be unable to amortise the costs of research and development over the long-term. Vendors often look for mechanisms to ensure lock-in; reducing the ability for clients to migrate to alternative platforms.

This has resulted in two different solutions:

- **Open Source Software (OS):** the source-code is available in an online and public repository under a liberal reuse license (such as the General Public License[8] and its affiliates); sometimes known as Free and Open Source Software (FOSS), not all open source software is free, and not all free software is open source; full customisation and extensibility is guaranteed;
- **Software-as-a-Service (SaaS):** software is available online on a centralised hosted server via a subscription service instead of as a deployable software system with a single, static price; upgrades, bug fixes and patches are consistently and regularly applied; custom extensions of functionality can be achieved via an API but customisation of the user interface is more limited;

Note that software can be both OS and SaaS, or neither.

Open source is also only useful if a developer community remains engaged and continues improving the software. The lifespan for open source software is often no different from proprietary solutions.

A combination of standards compliance and open APIs can ensure that you have the ability to migrate to a different service provider even where the software is not open source. If you must make a choice between standards compliance (i.e. offering a straightforward mechanism to migrate your data to another service) and open source, go for standards compliance.

Total cost of ownership for online software systems are often difficult to assess. Open Source products, where the software is effectively free, still come with a requirement for deployment, customisation and maintenance. A license fee for proprietary software is often a small part of a total lifetime cost.

Cost-of-life comparison between such vendors is difficult and NSOs are advised to present a clear brief on their needs to ensure that pricing is well-presented. This includes very specific guidance on data storage volumes and rate of growth.

4 Requirements and components for data publication systems

Open data systems were not originally designed to serve NSOs, nor were NSO systems originally designed for open data. While some data publication platforms offer **Extract, Transform and Load (ETL)**, Business Intelligence and Content Management functionality, this report concentrates on Data Discovery and Publication.

A visual representation of the various data and software components is presented as follows:

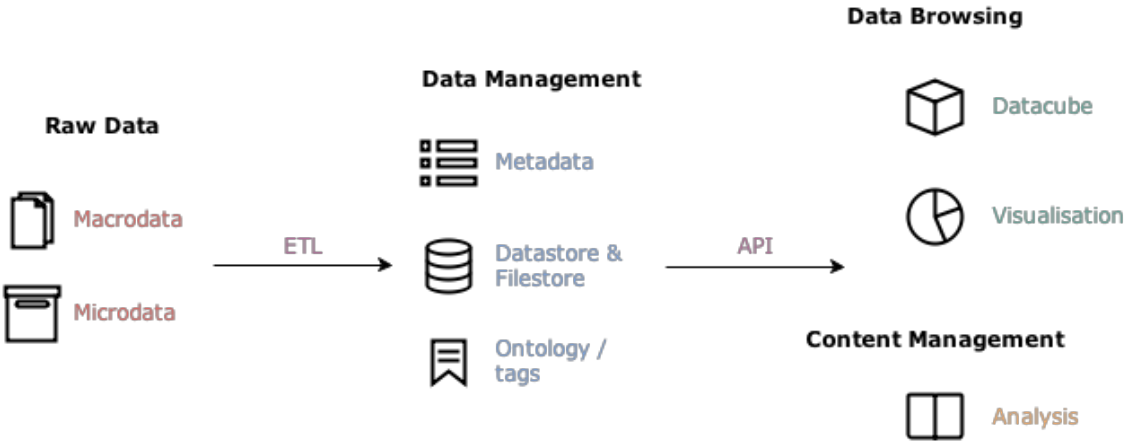


Figure 2: Overview of the technical components and processes in a data publication platform

The "Datastore & Filestore" component is the core of the service, providing both the data catalogue and presenting data in the raw file formats required for download. These components are often served by different software systems. Deployments should ensure system interoperability. Different software systems shouldn't result in duplicated functionality or needless requirements for manual data conversion for interoperability.

Table 1 presents assessment criteria used in this report and the relevance of each to open data and NSOs' traditional needs:

Components	Open Data Systems	NSO Systems
Descriptive metadata	●	•
Machine-readable	●	•
Anonymous access	●	•
Data licenses	●	
Data attribution	●	
Search	●	•
Open API	●	
Static URI	●	
Harvesting	●	
Federating	●	
Documentation	●	
Standards-based	●	
Structural metadata		●
OLAP hypercube		●
Data endpoints	•	●
Visualisation	•	●
UX & S/W extensibility	•	●

Table 1: Elements and components provided by data publishing software

Where ● is of primary importance, while • serves a secondary, or partial, role.

4.1 Criteria for Open Data

4.1.1 Descriptive metadata

Descriptive metadata are used for discovery and identification, as well as for data life-cycle management.

There are a large number of metadata vocabularies and ontologies used in open data systems and this is by no means an exhaustive list, but these are the most common:

1. **Data Catalog Vocabulary (DCAT)**[9]**:** an RDF vocabulary designed to facilitate interoperability of data catalogues published online and making extensive use of Dublin Core;
2. **Data Document Initiative (DDI)**[10]**:** an international standard for describing the complete research data life-cycle for the social, behavioural and economic sciences; the DDI-RDF Discovery Vocabulary is designed to support RDF;
3. **Dublin Core Metadata Element Set**[11]**:** an RDF vocabulary of fifteen "core" properties for use in resource description; elements include: title, creator, subject, description, publisher, date, format, source, etc.

Similarly, there are metadata systems for describing geospatial data, such as the **Infrastructure for Spatial Information in the European Community (INSPIRE)**[12] system which is derived from ISO 19115 "Geographic Information – Metadata". The revised version for 2014, "ISO 19115-1:2014 defines the schema required for describing geographic information and services by means of metadata. It provides information about the identification, the extent, the quality, the spatial and temporal aspects, the content, the spatial reference, the portrayal, distribution, and other properties of digital geographic data and services"[13].

As well as permitting data and metadata upload, user-friendly interfaces are required in which system administrators can associate metadata with each data resource.

4.1.2 Machine-readable datasets

Tim Berners-Lee summarises a recommended hierarchy of data availability[14]:

★	Available on the web (in any format) but with an open licence;
★★	Available as machine-readable structured data in a proprietary format (e.g. Excel instead of image scan of a table);
★★★	Available as machine-readable structured data but in a non-proprietary format (e.g. CSV instead of Excel);
★★★★	All the above plus standards from W3C (RDF and SPARQL) to identify data uniquely so that others can access and reference this live data;
★★★★★	All the above plus cross-link data to other data to provide context;

Spreadsheets and distributed data systems often lack an agreed data structure. A researcher who wishes to combine this with other data first needs to normalise it and then decide on standardised terms.

Converting semi-structured tabular data into a machine-readable format results in tabular files with a header row defining each of the data in the columns and rows below. Such tabular files (e.g. Excel) can easily be converted to a **comma-separated-value** (CSV) file.

Ignoring any further standards compliance, CSV files can be so arranged that they are "joined" on a common column. For example, a set standardised geospatial reference codes (e.g. ISO 3166 country codes) can be used to connect similar files covering different data series.

The process for converting data into a machine-readable format is one of **Extract, Transform and Load (ETL)**. A common requirement is parsing PDF files to extract tables into Excel or CSV and so render the data machine-readable. The process describes extracting data from its source, transforming it into the required machine-readable format, and loading it into a database or **Datastore**.

There is often a requirement to maintain a connection to the original data in case of a query or concern, and so data are often maintained in its original format in a **Filestore** alongside (and directly connected to) the Datastore.

4.1.3 Anonymous access

A key expectation of open data systems is that users can search for and access data and metadata without having to identify themselves, create a user account, or receive advance permission.

Experience to date strongly indicates that user registration schemes are significant deterrents to data access and use by broad audiences. When the UK Times introduced a paywall, they lost 66% of their readers overnight[15]. This happened despite the fact that users could register for free and access a limited number of articles each month.

Data managers often want to know two things about data reuse: who is using their data, and in what way? The former does not require user-registration, and the latter would not be addressed by user-registration.

In the case of analytical data, ensuring ease of access is critical to expanding the user-base beyond academics. The more approachable and user-friendly a data resource, the more likely that people will experiment.

Certainly, site managers need to understand data engagement: downloads, search terms, site referrals, and so on. None of this requires that users give up their anonymity. Similarly, no amount of user registration will tell site administrators how the data are used unless the user volunteers this.

Create opportunities for engagement, and make the process pleasant, and users will tell you themselves what they're doing with the data.

Note, this does not always imply that each dataset requires a full social media experience. Datasets are not opinion or blog-posts and comment threads attached to data are usually related to data quality. This requirement is best served by an issue tracker.

Data.gov.uk has an entire section dedicated to data-driven apps submitted by users. Some open data software systems permit users to not only create and save their own visualisations, but also feedback information on where those visualisations are embedded. That is useful feedback to the publisher and can still permit relatively anonymous engagement.

4.1.4 Data reuse and release licenses

Data publication software must offer a mechanism by which the license associated with each dataset is clearly presented to the user.

Licenses which permit data discovery but not liberal reuse are all but useless. If a person is not permitted to restructure or republish the data as they require then they are unlikely to want to use it at all.

Peter Desmet, a researcher at the Canadian Research Institute for Nature and Forest, describes how non-standard open access data licenses have made it illegal for him to aggregate 13,297 georeferenced American bullfrog records and place them on a single map[16]. This, despite the data being released as open access on the Global Biodiversity Information Facility (GBIF).

If an NSO is to meet its mandate for public dissemination, it must also ensure that the public has full rights to use that data in any form it may choose subject to due reference to the data publisher concerned. Datasets should be clearly labelled as released under standard open data licenses such:

- **Creative Commons (CC-By, CC-0)**[17],
- **Open Government Licence (OGL)**[18],
- **Open Database Licence (ODbL)**[19],
- **Open Intergovernmental Organisation License (IGO)**[20], or similar;

Each of these licenses permits the user to use the data they have downloaded, combine it with other data to create novel insight, and then to release or sell that data and insight as they wish subject – if required – to attribution to the original source.

4.1.5 Data attribution to source

Any dataset should present a clear set of information which permits a user to:

- link directly to the data;
- cite the data in their reuse of that data;
- attribute the data creator, either as an individual or as an organisation;
- contact a data owner should they have any queries;

Platforms must offer management and presentation of such data in a clear and readily presented format, as well as an interface to associate such attribution with source data. Where source data are not available, it is impossible for data users to offer appropriate attribution and for other users to verify the bona fides of the relevant data.

4.1.6 Search for data discovery

The default standard for online discovery is the natural language search box. Bing, Google and Yahoo are familiar examples. The process of data discovery offers stakeholders the ability to find relevant data quickly and easily, and then have access to that data in a useable format for the wide range of research activities which they may wish to perform.

Navigating via an impenetrable branching tree structure to find data - which has been structured according to the needs of the NSO and data publishers - is another barrier to data dissemination. Such structured systems can be useful for the expert user who has experience of that data structure, but – for everyone else – makes discovery extremely difficult.

The user needs the minimum number of steps to find appropriate data, verify that the data is what they're looking for, and then access that data in a format which permits them to use it.

Search results presented by data portals for their content should return focused summaries on datasets, along with keywords which aid classification, and the option of reviewing the data

online to assess its content. This may also include visualisation tools to produce basic charts or maps.

Faceting is encouraged when metadata are used as additional selection criteria to filter lengthy search results. A simple mechanism is for metadata to be listed alongside search results with check-boxes that automatically apply an “and” to filter the results.

4.1.7 Application Programming Interfaces (APIs) are public

Interoperability requires an **Application Programming Interface (API)** through which standardised commands are available to an external system and used to query the data, metadata and other attributes in a database.

Such interoperability permits a range of actions by other software systems, for example:

- Import data into another application (such as Tableau, R, or Excel) for analysis and merging with other data sources;
- Development of free-standing applications, such as transport apps on mobile phones;
- Automate repetitive processes, such as setting a routine to regularly download data released monthly;

APIs can also be used by site administrators to automate data harvesting, uploading or similar bulk processes. Such APIs are often connect to ETL systems for data transformation prior to loading.

The most common approach to implementing an API is via a **Representational State Transfer (REST)** system. The standard commands generally used in REST for creating, reading, updating, or deleting data are POST, GET, PUT, DELETE.

The more sophisticated software platforms often have a query-builder utility which permits users to experiment live on the server and see the results of different API queries. Such interfaces should permit unique URLs so that particular queries can be bookmarked and shared.

4.1.8 Datasets are reachable via persistent URI

Permanent and persistent availability means not just that data are available, but also that they are always in the same place.

For online systems, this implies a requirement for **Uniform Resource Identifiers (URIs)** which ensure that resources, content or data are always located through one discrete address for any and all users or software-driven applications. The most familiar of these are the **Uniform Resource Locators (URLs)** that you see as links to websites. These can also be described as endpoints.

It is essential that these endpoints never change and that their behaviour is predictable. A person who bookmarks a link, or who includes such a link in an article, assumes that it will still be there when they need this.

More generally, a URI can also define a persistent link to a point within a dataset. This permits the interlinking of different datasets, the creation of more complex data aggregations and better insight into that data.

Software should be capable of providing a clear and easy to find permanent URLs for every dataset being served by the platform. Some systems are capable of providing URIs to subsets of data, or points, within datasets.

4.1.9 Automated data harvesting

The ETL process of capturing data for publication can often become a significant barrier to data migration or implementing a new data system.

The manual process of creating datasets, entering metadata, and uploading data, is slow and labour-intensive. It is also difficult to automate since identifying and allocating appropriate metadata is a specialist task often difficult to extract from the data files.

Some data, however, are updated regularly as part of a data release cycle. Datasets already exist and merely need to be updated.

Where software has an API which permits data editing and uploading, custom scripts can be written which will automate such processes. Some software systems go further, offering a dashboard to system administrators allowing them to set up and manage a large number of automated processes for data upload. Such as service is known as data **harvesting**.

Harvesting is the data publication receiving end of an ETL process which is usually delivered by alternative software systems.

The more such routine tasks can be automated, the easier data release becomes and the more likely that both users and providers of data will adopt the system.

4.1.10 Federation of multiple data sites

Federation is the mechanism by which dataset metadata are polled from different platforms and copied to a centralised software service or database. The original data usually continues to be hosted on the original platform but the metadata, and links to the data resources, are now accessible and discoverable via the platform's search engine.

There are numerous reasons why data may be published from a variety of different software platforms. Different departments may wish to manage their own data life-cycle. Federal agencies and ministries often enjoy significant autonomy and they will be used to operating their own systems.

Open data, however, is more useful when users do not need to visit numerous different web services in order to discover data.

Note that federation differs from harvesting. Data which is harvested is maintained and managed from that system. Federated data captures only the metadata but also permits live exploration or visualisation of that data in the remote system.

Not all software which permits harvesting will support federation.

As with harvesting, an API which permits search and discovery also permits federation. A system can traverse the data for the site and build a metadata structure for local search. More sophisticated software systems offer dashboards for automated federation.

Federating between unrelated software platform offers challenges as a result of different database schema. Same software federation is often straightforward while specific software adapters (transformations) are required to federate across unrelated software systems.

Federated systems need to regularly poll their data sources to ensure continued alignment in the case of deletions, updates or additions.

4.1.11 Public documentation

APIs, faceting, data reuse are - by their very nature - technical topics. Different platforms behave differently and even experienced analysts will struggle in the absence of clear documentation.

Software documentation is the responsibility of the platform vendor and it is essential that they provide comprehensive information for developers and the general public on how their platform works. Such documentation should be updated with each new software release.

For developers, demonstration services which encourage experimentation with the use of APIs are extremely helpful.

It is also helpful if such data are available online instead of as PDF documents. This permits search, cross-referencing and persistent URIs. Machine-readability is as important for documentation as for data.

4.1.12 Compliance with generally accepted standards

The definitions of open data systems are still emerging as best-practice is agreed. The leading open data software systems tend to have similar approaches to metadata and in data presentation. Many of these are becoming standards recognised by the W3C especially as regards metadata, RDF, and hypercubes.

The list of components presented in this section, 4.1, are a reflection of the current generally accepted standards. Software which deviates from this in any significant way is more likely struggle with interoperability and general compliance.

4.2 Criteria for National Statistics Offices data publication

4.2.1 Structural metadata

Even in the unlikely situation where an NSO has implemented an integrated software platform for all their data publishing needs, there is still a requirement for researchers to access that data and use it in their own systems.

The internet offers the ability to connect and mash together a wide variety of data in different formats to produce new insight. Writing in 2001, Tim Berners-Lee pointed out that the majority of information on the web is designed for people, rather than computers, to read[21].

Gareth McGuinness, of the International Monetary Fund, describes the challenge: “For each dataset, the IMF must go through the laborious work of matching each dimension in the source data to the equivalent IMF dimension, and then matching each item in each code list to the equivalent item in the IMF code list. In the best case scenario, there will be some instance where code lists match. However, even for one of the simplest geographic dimensions – country – there are several different “standards” in common use.”[22]

Semantic interoperability is the ability for computer systems to exchange data unambiguously. Structural metadata permits alignment of the data itself to perform restructuring or use in linking to other, similar, datasets.

A number of metadata formats are used by NSOs to structure or promote data interoperability, and these are the leading vocabularies:

1. **PC-Axis**[23]**:** is a software suite developed by Statistics Sweden – and in use by more than 50 NSOs around the world – which provides a set of structured keywords defining the file format for loading data as a cube; it was initially developed in the 1980s for use with the Axis database system but has been extended;
2. **Statistical Data and Metadata eXchange (SDMX)**[24]**:** a mechanism for the exchange of statistical information. The initiative is sponsored by EUROSTAT, IMF, OECD, UN and World Bank, amongst others. This is an extremely detailed approach which offers a language in which different statistical data can be integrated across different software systems. SDMX creates a mechanism for mapping existing NSO metadata to SDMX and forming a hypercube for custom slices to be extracted.
3. **RDF Data Cube Vocabulary**[25]**:** provides a means to publish multi-dimensional data using RDF and compatible with SDMX; the RDF Data Cube vocabulary is a core foundation which supports extension vocabularies to enable publication of other aspects of statistical data flows or other multi-dimensional data sets;

DDI, described in 4.1.1, is also used by NSOs to define structural metadata. The RDF approach to interoperability is the **OWL Web Ontology Language**[26].

4.2.2 OLAP hypercubes

Conforming machine-readable data into a relational database results in an array of data in multiple dimensions; an online analytical processing cube, or **OLAP cube**. Such data are now available for integrated analysis and straightforward software-driven conversion into multiple formats.

The term “cube” can be misleading as an OLAP is not limited to only three dimensions. The UK government financial data from Combined Online Information System (COINS) was converted into a linked data system in June 2010[27]. Each datum in COINS is uniquely identified from a combination of seven indices in a structure called a **hypercube**.

A cube consists of as many dimensions as required to define its data uniquely. The terms OLAP, datacube and hypercube will be used interchangeably in this report.

Getting from individual spreadsheets to a neatly aligned, comprehensively analytical, datacube requires a process of **Data Structure Governance** – agreeing on data structures across entire organisations – as well as ETL.

Hypercubes permit filtering and faceting of the data itself. Users can select particular series, for a range of geographies, and over specific dates, to create a custom data slice.

Statistical data are used to develop research insight. Individual tables become more useful when they can be aggregated together and sliced into numerous views for analysis.

4.2.3 Data endpoints

Structured data endpoints return data in predictable ways. These can be as simple as a known type of serialisation format while more complex implementations permit the data to be queried, filtering or refining the dataset prior to download.

What they have in common is that there is a fixed URL to reach the end-point.

A number of commonly-used endpoints are:

1. **JavaScript Object Notation (JSON)**[28]**:** an open standard presenting data as a set of key-value pairs; as an end-point, it is accessible via RESTful APIs;

2. **OData**[29]: While RDF has achieved a high degree of traction for linking diverse data together, it is still not that straightforward to connect it back to the tools researchers use most frequently to work with data. OData offers a standardised protocol for creating and consuming data. The format is extremely popular, and software as diverse as Tableau (for analysis and visualisation), Drupal (for content management), and Microsoft's Excel are all able to accept OData as an input. OECD.stat, for example, is currently offering a test interface to trial OData[30].
3. **Extensible Markup Language (XML)**[31]**:** a markup language, more difficult for humans to read, found in diverse uses such as encoding Microsoft Office documents, websites and for data interchange; it too can be presented via RESTful APIs;
4. **SPARQL Protocol and RDF Query Language (SPARQL)**[32]**:** an RDF database query language able to retrieve and manipulate RDF data, returning it as RDF or – with appropriate interpreters – conversion for use by SQL or other query languages;

4.2.4 Online analysis and visualisation

Business Intelligence systems offer meaningful tools to stakeholders who wish to perform statistical analysis across the complete data produced or managed by an NSO. Such tools may be as simple as selecting a data slice and displaying it as a chart, or performing complex statistical analysis on multiple data series, as well as presenting on maps and charts.

User engagement in the face of vast and complex data is more likely where that data are presented in a tactile and engaging way. The growth of data journalism has depended on the growing availability of data resources and the presentation of exciting views on that data.

Certainly, NSOs should not editorialise or lose their neutrality, but they can provide tools for users to explore the data. Technical researchers and analysts will want to import that data into their favourite systems and data endpoints will permit them to do so.

Whether a software platform provides its own native visualisation and business intelligence tools is often a design decision. Some systems prefer an integrated approach, while others prefer to focus only on data publication. Which approach will serve better depends on any existing infrastructure in the deployment environment, or on the needs and expectations of the NSOs.

4.2.5 User-experience and software customisation

Data publication platforms are only one software system amongst many deployed by NSOs. Software should permit sufficient template and layout customisation to provide a consistent user-experience and provide a common look and feel across all NSO online services.

Sub-domains, which are the URLs to different software in a suite of platforms, are part of providing a consistent user-experience. If NSOs have chosen a particular addressing format (e.g. reports.nso.gov, data.nso.gov, events.nso.gov, etc.), it is important to verify that this is possible where software is provided as SaaS, should the NSO require.

Similarly to user-experience customisation, NSOs often require custom software extensions for integration with existing platforms, or to perform particular tasks.

NSOs may choose to develop custom software extensions and enhancements in-house or via third parties.

Where software is licensed as open source, then such development is relatively straightforward (with due consideration to documentation mentioned in 4.1.11). If the software is not open source, it is essential that a well-documented API permit software extension, or that licensees have access to the source-code in the absence of APIs.

In the case of open source, or licensed code access, the programming language (Python, Java, Ruby, C++, etc.) becomes a consideration. APIs, as mentioned before, are software agnostic and permit development in the programming language of your choice.

5 Review of Open Data Publication Systems

This section presents a review of a selection of software identified as in-use by NSOs or currently considered to be viable open data platforms.

During the research stage for this report, custom software implementations were also assessed. Organisations like the US Census Bureau have developed their own system. In large part such work was initiated long before there was an open data movement, let alone software to support it. Such custom software can often become more generally accepted amongst closely-related NSOs and be released more formally. This is true in the case of Nesstar and PC-Axis.

By no means is this an exhaustive list and it is likely that this comparison will be updated from time-to-time.

Software platform	Descriptive Metadata	Machine-readable	Anonymous access	Data licences	Data attribution	Search	Open API	Static URI	Harvesting	Federating	Public Documentation	Standards-based	Structural Metadata	OLAP Hypercubes	Data Endpoints	Visualisation	UX & S/W Extensibility
CKAN	●	●	●	●	●	●	●	●	•	•	●	●			•	•	●
DevInfo	●	●	●	●	●	●	●	●			●		●	●	•	●	•
DKAN	●	●	●	●	●	●	●	●	•	•	●	●			•	•	●
Junar	●	●	●	●	●	●	●	●	•	•	●	●			•	●	●
NADA	●		•	●	●	●		●			•	●	●				●
Nesstar	●	●	●	•	•	●		●			•			●		●	•
OpenDataSoft	●	●	●	●	●	●	●	●	•		●	●			●	●	●
PC-Axis and PX-Web	●	●	●	•	•	•	●	●	•	•	•	•	●	●	●	●	●
Prognoz	●	●	●	•	•	●		●						●		●	•
Semantic MediaWiki	●	●	●	●	●	●	●	●			●	●			•		●
Socrata	●	●	●	●	●	●	●	●	•	•	●	●			●	●	●
Swirrl	●	●	●	●	●		●	●	•	•	●	●			•		●

Table 2: Software commonly found supporting online data publication

● offers a complete solution
• offers a partial, or incomplete, solution

Rows highlighted in blue are for software commonly regarded as open data platforms. Note that implementations of various elements may differ even as they offer similar functionality.

5.1 CKAN

URL	http://www.ckan.org/
S/W Licence	Affero GPL, open source
Language	Python, Javascript
SaaS	http://ckanexpress.com/
Demo	http://demo.ckan.org/
Examples	data.gov.uk
	data.surrey.ca
	publicdata.eu

Overview

CKAN is an open-source, data discovery system making data accessible and usable by streamlining publishing, sharing, finding and using data. As well as harvesting, cataloguing, and advanced searching, it can store data and provides rich data APIs, and simple visualization and exploration tools.

Most CKAN instances are self-deployed and self-hosted by the organisations concerned, although a number of companies now offer CKAN as SaaS. There are also a globally distributed range of consulting and software services available to deploy CKAN.

Open data suitability

Descriptive metadata	CKAN has the concept of the dataset being a folder in which files, known as resources, are stored. Metadata are applied at the dataset level and not at the data structure level. Metadata is served as RDF and support Dublin Core, and DCAT, with the ability to harvest documents in the geospatial INSPIRE format.
Machine-readable	CKAN is able to import and interpret CSV, XLS, GeoJSON and text files in a machine-readable format. It is further able to interpret and serve PDFs and other file-types even where it cannot import this into the database. Geospatial support includes ArcGIS through extensions, and the ability to harvest INSPIRE and other ISO19139 based geospatial metadata.
Anonymous access	Anonymous access is permitted, although CKAN does have the ability to offer private datasets to subsets of registered users. An API key is required for users wishing to modify or upload data via the API.
Data licenses	CKAN offers a range of common license types to the user during the data upload process and these are presented clearly in the dataset view.
Data attribution	CKAN presents clear attribution where such data exists.
Search	CKAN's search is clearly presented, and returns results permitting filtering by metadata. Search results are presented with title and description, and a list of file-types present in the dataset.
Open API	CKAN has a clear, although complex, API that is documented thoroughly online as part of the main documentation. Data resources can be downloaded via the API but are only available as JSON output from the database.
Static URI	CKAN presents all datasets and data resources as static URIs.
Harvesting	CKAN is able to harvest existing data resources, as well as regularly changing data, via the API. There is also a data harvester extension which permits a limited user interface for setting up individual harvesting processes.
Federating	CKAN is able to federate other CKAN sites in order to consolidate metadata. The harvester extension provides a common framework for developing custom harvesters for

	different metadata sources, which means that federation from non-CKAN sources is also possible. Some examples include generic spatial metadata sources like CSW and WAF, and ArcGIS Server portals.
Documentation	CKAN's documentation is comprehensive and available at http://docs.ckan.org/. This presents both information for users as well as for developers.
Standards-based	CKAN is an entirely open source software platform, although the learning curve for the original Pylons framework underlying CKAN can be steep. Extending CKAN is performed via either of the software library (in which case the extensions run on the server with CKAN) or via the API (where they can then run remotely).

NSO suitability

Structural metadata	CKAN offers no support for metadata for data structure.
OLAP hypercube	CKAN places greater emphasis on the source data than on the database. There is no support for hypercubes. Data can be filtered or faceted via the API, although such support will require custom software development to be generally useful.
Data endpoints	CKAN produces JSON output, and – through an extension – also supports OData. CKAN only presents data in the format in which it was uploaded and does not offer data transformations. CKAN endpoints do not support any proprietary formats like Stata or SPSS, although extensions can be developed.
Visualisation	CKAN's included visualisation library, Recline.js, is very limited in terms of what it can support. It does not permit the user to save or share specific visualisations either, although this is planned for a future version. A new Dashboard feature offers a range of persistent visualisations to the site administrators but not to visitors.
UX & S/W extensibility	CKAN uses Bootstrap, a popular UX library, as its CSS system, with Jinja2 for templating, and so can be easily customised. It should be noted, though, that CKAN is not responsive in design and does not support mobile screen form factors. While CKAN predominates in serving national open data portals, its real strength is in the community of developers producing custom extensions and enhancing the software. The UK, US, Canada, Australia and even Mexico governments have each supported local developers. Universities have also acted to extend CKAN to host academic research data.

Observations

CKAN is a powerful, and extensible, platform for open data systems but will need hypercube, and data structure metadata, support before it can be deployed for integrated NSO statistical portals.

5.2 DevInfo

URL	http://www.devinfo.org/
S/W Licence	Free but license unclear
Language	Unknown
SaaS	http://www.devinfo.org/
Demo	http://www.devinfo.org/
Examples	All sites hosted on http://www.devinfo.org/

Overview

DevInfo was designed to integrate the data generated through monitoring the Millennium Development Goals and is developed and supported by UNICEF. Numerous UN agencies use variations of DevInfo and governments around the world are encouraged to publish their MDG indicators in it.

The software is available free and can be deployed on your own servers. This gives you the opportunity to use what is quite a powerful business intelligence tool to manage and present your own data.

Open data suitability

Descriptive metadata	Metadata is entered via a specific module in the software and is compliant with DDI and Dublin Core, as well as ISO 19115 for geospatial data.
Machine-readable	DevInfo's objective is to align all data into a datacube. Only machine-readable data, formatted according to the specific platform requirements, can be read by the system. It does not host a variety of data objects.
Anonymous access	Anonymous access is a default, and logins are available for site administration. Registration and login is required to download data or save visualisations.
Data licenses	There is no mechanism for data contributors to declare data reuse licenses and it is unclear what those licenses may be.
Data attribution	Links and references are provided for each data-series for appropriate attribution, including – where available – a reference email for direct contact.
Search	Search is limited to exact matching of location and data series. Natural language and metadata search are extremely limited.
Open API	An API is provided, including an online tool for experimentation with generative calls. The request format supports both REST and SOAP, with output in XML and JSON, as well as SDMX.
Static URI	There are no static URIs available for data series. DevInfo appears to be structured as a "one-page website" with no URL changes despite different views being created. Once data has been selected and saved then it is possible to generate a static url, or even to embed the data, but the approach of search-and-link is absent.
Harvesting	Bulk uploading is possible through additional DevInfo modules but establishing regular, automated data harvesting is not available via an interface. This could be written as a script, however.
Federating	While it is possible to search all sites hosted by DevInfo on the main site, there is no automated mechanism for federating remote sites hosted independently.
Public Documentation	DevInfo is not open source and the documentation is only that provided by the UNICEF team. All the documentation is only available in zipped and PDFed form making using it

	somewhat tedious. Documentation is aimed at end-users rather than at developers. All content is available from the main DevInfo website.
Standards-based	The software stack runs only on Windows Server and Microsoft SQL Server. This means that it will need entirely separate architecture from any open source software the NSO may choose to deploy.

NSO suitability

Structural metadata	DevInfo supports SDMX for data output and via the API, but structures data via its own format internally.
OLAP hypercube	Data is stored in a hypercube but the interface for creating custom slices of that data is limited. Once the user has selected data – with the ability to select multiple data series – the user then has the option to slice that data by time-period as part of the data visualisation component.
Data endpoints	Endpoints are available via the API or once a user logs in. This does, however, limit a user's ability to integrate data with their own applications.
Visualisation	The software offers a clear, user-friendly data browser and business intelligence tool which should suit most online users. The user is able to create a range of standard charts, including plotting geospatially, and to name axes and change the colour presentation of the data.
UX & S/W extensibility	The software is not open source and there is limited documentation. Beyond changing a few logos, there is almost no way to customise the user-interface and software extensibility is extremely limited. The lack of a development community also means that any NSO would have limited support for customisation.

Observations

DevInfo offers a comprehensive data management and publication platform. It does not meet all the criteria required for open data publication, especially regarding anonymous access, and data licensing. DevInfo does appear to serve the immediate needs of NSOs.

That said, DevInfo is not very customisable and NSOs with requirements which vary from these will find it difficult to modify the system. It was designed to serve Millennium Development Goal data. Less specific data series may fare less well.

5.3 DKAN

URL	http://www.nucivic.com/
S/W Licence	Open source, GNU GPL
Language	PHP, JavaScript
SaaS	http://nucivic.com/data
Demo	http://demo.getdkan.com/
Examples	whitehouse.gov/raise-the-wage
	abrepr.org
	www.offenedaten-koeln.de

Overview

DKAN is an open source solution built on Drupal, a leading content management system for thousands of governments worldwide, and aligned with the data standards and best practices of the CKAN data portal software. Like CKAN, DKAN is a data discovery system making data accessible and usable by streamlining publishing, sharing, finding and using data. As well as harvesting, cataloguing, and advanced searching, it can store data and provides rich data APIs, visualization and exploration tools.

Additionally, unlike CKAN, DKAN is a distribution (pre-configuration) of Drupal and as such is also a complete CMS offering comprehensive tools to manage content, documents, and community, in addition to data sets. This also gives DKAN access to the tens of thousands of Drupal developers and extensions already developed for the platform. Key CMS features include blogs, groups, taxonomies, WYSIWYG editing, faceted search, form building, calendars, and a full graphical user interface for administering all content, workflows, user roles and permissions.

Open data suitability

Descriptive metadata	DKAN, similarly to CKAN, has the concept of the dataset being a folder in which files, known as resources, are stored. Metadata are applied at the dataset level and not at the data structure level. Metadata is served as RDF and support Dublin Core, DCAT and the INSPIRE geospatial format. Drupal supports the creation of custom metadata as well.
Machine-readable	DKAN is able to import and interpret CSV, XLS, XLSX, and text files in a machine-readable format. It is further able to interpret and serve PDFs and other file-types even where it cannot import this into the database.
Anonymous access	DKAN is able to use the full Drupal authorisation system, including permitting anonymous access to public datasets for search and download.
Data licenses	DKAN offers a range of common license types to the user during the data upload process and these are presented clearly in the dataset view.
Data attribution	DKAN presents clear attribution where such data exists.
Search	DKAN's search is clearly presented, and returns results permitting filtering by metadata. Search results are presented with title and description, and a list of file-types present in the dataset.
Open API	DKAN has a clear, although complex, API that is documented thoroughly online as part of the main documentation. Data resources can be downloaded via the API and available as JSON or XML output. An API key is required for users wishing to modify or upload data via the API. No API key is required for search or download.
Static URI	DKAN presents all datasets and data resources as static URIs.
Harvesting	DKAN is able to harvest existing data resources, as well as regularly changing data, via the API. There is currently no user-interface for setting up automated harvesting tasks,

	however, it should be possible to use the CKAN harvester for this.
Federating	Drupal is able to federate with multiple Drupal sites and so, intrinsically, this is possible with DKAN. However, this has not been tested to any large degree.
Public Documentation	DKAN's documentation is comprehensive and available at http://docs.getdkan.com/. This presents both information for users as well as for developers.
Standards-based	DKAN is aligned with best practice in the open data industry.

NSO suitability

Structural metadata	DKAN offers no support for metadata for data structure.
OLAP hypercube	DKAN places greater emphasis on the source data than on the database. There is no support for hypercubes. Data cannot be filtered or faceted via the API, although support for OData may permit this in future.
Data endpoints	DKAN produces JSON and XML output. DKAN presents data in the format in which it was uploaded and does not offer data transformations.
Visualisation	DKAN's included public facing visualisation library, Recline.js, is very limited in terms of what it can support. It does not permit the user to save or share specific visualisations either. Recently, DKAN has developed a more sophisticated visualisation system for embedding and saving charts, including geospatial data, as part of data-driven story-telling. While not a sophisticated business intelligence tool, this does provide entry-level data presentation services. Overall, rather than trying to be a robust data visualisation tool itself, DKAN has developed a toolkit to facilitate integration with third-party data visualisation web services such as CartoDB.
UX & S/W extensibility	As a set of Drupal components, DKAN also has the advantage of being part of one of the most active open source projects in the world. The range of services and software available to Drupal is extensive, including thousands of skilled developers available internationally. Drupal is one of the leading content management systems and is used by many of the world's most popular websites. The user-interface is based on leading best-practice, and the software is extremely extensible. DKAN has developed flexible UX tools to map its schema (and that of any Drupal content type) to any other schema, such as CKAN or US Project Open Data.

Observations

DKAN is a powerful, and extensible, open source platform for open data systems, including CMS support. There is also the option of enterprise-level SaaS. However, it will need hypercube, and data structure metadata, support before it can be deployed for integrated NSO statistical portals.

5.4 Junar

URL	http://www.junar.com/
S/W Licence	Proprietary
Language	Java, Django / Python
SaaS	http://www.junar.com/
Demo	http://www.junar.com/
Examples	data.sanjoseca.gov
	datosabiertos.gob.go.cr
	data.cityofsacramento.org

Overview

Junar is a specifically Software-as-a-service platform offering one of the leading open data platforms. The system is able to import and use a wide variety of data formats and, as with all SaaS offerings, is useful to users looking for rapid deployment and the ability to develop and present insight from their data very rapidly.

Junar has also developed a proposition for academic publishing and so is extending into new markets.

Open data suitability

Descriptive metadata	Junar uses RDF metadata to describe the datasets, presented in Dublin Core and DCAT.
Machine-readable	Junar offers a wide range of support for different machine-readable formats, including CSV, XLS and XLSX, JSON and SOAP/XML 2.0, as well as the KML, KMZ, GeoJSON, and Shapefile geospatial formats.
Anonymous access	Users do not need to create accounts in order to access data, but they will need to do so in order to access higher services.
Data licenses	Licenses are not clearly presented for individual datasets. The next software release will include custom licences for datasets using the template provided by http://project-open-data.github.io/license-examples/
Data attribution	Additional info for each dataset provides links to the source data.
Search	Junar offers search, but this is not – visually – a priority on the site. One weakness is that Junar has limited focus on faceting data. Data may be structured with metadata but that is not exposed through the interface to the user to permit them to filter search results in a more accessible way.
Open API	Junar provides an interactive API for each of their sites. This permits developers to experiment live on the database to see what results they can achieve.
Static URI	Each dataset generates a unique URI. Visualisations and dashboards created by the system administrator also get a unique URI.
Harvesting	Junar has an additional component called Publishing Workflow which permits some automated collection and management of data from different locations, including assigning metadata to that data. The Junar Uploader is a set of scripts for automatically uploading a range of file-types, including CSV, XLS, XLSX, KML and KMZ. Junar has integrated capabilities to collect data from REST/JSON or SOAP/XML web services linked directly to source databases for real-time, or near-real-time, data collection. Integrated REDATAM+SP software permits harvesting from HTML forms to collect data directly.
Federating	All Junar sites run as SaaS from the same servers but, at this

	stage, data are not federated across the different platforms. Junar has produced an extension for CKAN so that CKAN is able to read metadata from Junar and present it in search results.
Public Documentation	The documentation for Junar is available on a set of wiki pages, many of which are customised for particular clients http://en.wiki.junar.com/index.php/Main_Page. This is not particularly easy to read and of limited value to developers. Given that this is a critical part of the service, this needs to be much more visible, both from the Junar main site (which would benefit from an entire Developers subsite) and from the client sites. There is a Knowledge Base available at http://support.junar.com which is not regularly updated but contains a basic subset of information regarding the use of the platform, the Publishing Workflow, an FAQ section and a Feature Request section.
Standards-based	Junar has focused on supporting the leading standards in the open data community.

NSO suitability

Structural metadata	Junar does not currently support structural metadata.
OLAP hypercube	There is no hypercube support.
Data endpoints	Junar supports a wide range of data endpoints, including CSV, JSON, PDF, RDF, RSS, XLS, XLSX, XML, all of which are also available via the API. They have also recently added OData to their list, giving the opportunity to query individual datasets as well. They also provide integration with Google Docs and Dropbox.
Visualisation	Junar's main focus is on providing support for the development of comprehensive data-driven visual dashboards. While discovery is important, Junar has realised that many of their clients also want to use their data to tell stories which offer stakeholders an easy way to digest complex data. The standard range of charts are available, including geospatial plotting, and the ability to drag and drop a wide variety of visual types to create an integrated dashboard.
UX & S/W extensibility	Junar is proprietary software and the range of published public APIs are only about downloading data rather than extending functionality or uploading data. While the user-interface can be customised, and new functionality written, the NSO is reliant on Junar for these services.

Observations

While Junar is one of the leading open data publication services, it has no support for structural metadata or hypercubes required by NSOs.

Junar, as with the other proprietary SaaS services, concentrates on ease of deployment and providing visual tools with plenty of hooks for downloading and developing custom applications. From a client perspective, this is a straightforward approach to getting open data to the public quickly.

5.5 NADA

URL	ihsn.org/home/software/nada
S/W Licence	Open Source, BSD
Language	PHP
SaaS	
Demo	
Examples	microdata.statistics.gov.rw
	statistics.knbs.or.ke
	nigerianstat.gov.ng/nada

Overview

NADA is a web-based cataloging system that serves as a portal for researchers to browse, search, compare, apply for access, and download relevant census or survey information. It was originally developed to support the establishment of national survey data archives. The application is used by a diverse and growing number of national, regional, and international organisations.

While the platform is open source, the additional DDI Metadata Editor is proprietary and provided by Nesstar Publisher as freeware.

The International Household Survey Network coordinated by the World Bank responsible for maintaining NADA have discussed a thorough rearchitecture of the platform onto the Symphony developer framework. This is what Drupal, the CMS, is developed on and promises a future in which NADA integrates well and is more readily deployed. It is currently on the CodeIgnitor Framework.

Open data suitability

Descriptive metadata	All resources are associated with DDI metadata which is not produced from NADA itself. The DDI Metadata Editor produces DDI compliant XML files for upload into NADA. It is not an obligatory as part of the platform – there are other editors and advanced users can even use a text editor – but this one happens to be free. NADA also presents the metadata in RDF.
Machine-readable	NADA is specifically designed for managing and presenting microdata. It does not have mechanisms for interpreting data resources as machine-readable. Any and all data resources are stored and presented as-is for download.
Anonymous access	Anonymous access for search is default, however – as this is designed for microdata - access for data download requires a login and appropriate authorisation. If the direct access, or recently added open data license types, are chosen then no login is required. Users then only agree to terms appropriate for the license and go directly to download the data files.
Data licenses	Each dataset contains comprehensive details on licensing and reuse.
Data attribution	Each dataset contains comprehensive details for data attribution. Importantly, NADA offers clear citation references as part of an international drive to encourage better citation of data and recognise it as citable work.
Search	Full text search is provided, with filtering and faceting, including the ability to limit the search space by the data range of research publication.
Open API	NADA does not have a published API which makes extension more difficult. There is a private and undocumented API which offers access to a few of the DDI metadata fields. The lack of API or mechanism to manipulate data means that aggregations cannot be derived from the microdata programmatically.

Static URI	Every view presents its own URL, including the resources for download.
Harvesting	NADA does not support harvesting, although – given that DDI can be captured in an XML file for import – a mechanism for developing such automated import should be feasible.
Federating	NADA does not support federation.
Documentation	Current documentation is available in PDF and on documentation.ihsn.org/nada/4.2/. Overall developer documentation is limited and until the new Symphony framework is adopted – potentially more than 12 months away – this will continue to be somewhat forbidding for custom extension development.
Standards-based	NADA is designed to support the exacting requirements for DDI metadata and document lifecycle management. It is standards based.

NSO suitability

Structural metadata	NADA supports DDI structural metadata requirements.
OLAP hypercube	There is no hypercube support.
Data endpoints	Metadata are available as XML and the description of the external resources are available in Dublin Core, but the data itself are only available as the originally uploaded source document.
Visualisation	NADA provides no support for data visualisation.
UX & S/W extensibility	The developer documentation is limited but NADA is open source and so customisation and extension is possible with some trial-and-error. Once the system is ported to Symphony, customisation should be far easier.

Observations

NADA is the only platform in this survey which is designed expressly to support the needs of producers and archives publishing microdata. Most users are National Data Archives, some universities and some international organisations. It is not designed as an open data platform and does not provide an API or interpret machine-readable resources. It also does not support the integrated needs of NSOs.

5.6 Nesstar

URL	http://www.nesstar.com/
S/W Licence	Proprietary
Language	Java
SaaS	
Demo	http://nesstar-demo.nsd.uib.no/
Examples	nesstar.ess.nsd.uib.no
	nesstar.ukdataservice.ac.uk
	nesstar.ssc.wisc.edu

Overview

Nesstar offers a vertically integrated suite of tools for data publishing and management. Nesstar Publisher consists of data and metadata conversion and editing tools, enabling the user to prepare these materials for publication to a Nesstar Server. However, it can also be used as a stand-alone tool for the preparation of data and metadata. The Publisher enables users to enhance datasets by combining a wide range of catalogue and contextual information, which can then be viewed within the Nesstar web client, Nesstar WebView.

Nesstar offers support for multilingual metadata, microdata, aggregate data, multi-layered maps, various visualization, subscriptions/notifications, cell notes/missing data symbols, basic analysis and embedding of live data into regular web pages.

Open data suitability

Descriptive metadata	Nesstar supports both Dublin Core and DDI for descriptive metadata.
Machine-readable	A very large range of data files are acceptable as data input, including: NSDstat, DDI, SPSS, Stata, Statistica, dBase, DIF, CSV, PC-Axis, Excel and Hierarchy Definition Files. This is amongst the most comprehensive of data format systems. Where non-machine-readable files are imported, such as PDF or Microsoft Word documents, Dublin Core and e-GMS are used to define the descriptive metadata. Geospatial data is supported through deployment of GeoServer.
Anonymous access	Anonymous access is available for aggregate (i.e. generalised data) but not automatically for the microdata stored in the system. An authentication system controls access to such data unless specifically set as direct download.
Data licenses	Additional variable data and links to licenses can be supplied along with data descriptions or associated metadata files, but there is no standardised mechanism for listing and declaring data licenses.
Data attribution	A descriptive metadata file is provided with each data-series, offering the complete reference for the data.
Search	Search is rudimentary, returning results as a branching-tree with little guidance as to where the results may be found. The main mechanism for search is a tree list of all the data resources. There is a more comprehensive advanced search, but this will require some knowledge of the data the user is hoping to find.
Open API	A RESTful API is provided for the platform, including comprehensive documentation on a Git repository at gitlab.nsd.uib.no/nesstar/nesstar-rest-api/. There is no online interactive demonstration of the API.
Static URI	While not immediately obvious, clicking on the link icon in the data browser does provide a static link to each of the data.
Harvesting	Nesstar now supports the Open Archives Initiative Protocol for Metadata Harvesting[33]. A new standalone component

	allows server administrators to expose a server's metadata for harvesting by others. OAI-PMH is a standard protocol designed to make it simpler for data providers to open up their repositories and for service providers to harvest metadata. The protocol uses XML over HTTP and supports Dublin Core and DDI.
Federating	Nesstar does not appear to be designed for federation.
Public Documentation	Documentation is fairly comprehensive and, since the software is designed for self-deployment, system administration and customisation documentation also exists. Note, however, that much of the documentation is only available as PDFs from their site. Each of the products, Server, WebView, and Publisher, are presented there. Critically, the Server documentation is available online and is searchable.
Standards-based	While Nesstar is proprietary, the extent of support for metadata standards is comprehensive. Additionally, while proprietary, Nesstar is developed using the open source JBoss suite of middleware. Nesstar can be installed on Windows or Linux systems.

NSO suitability

Structural metadata	DDI is used for structural metadata.
OLAP hypercube	Nesstar has comprehensive datacube support, permitting files to be imported directly into the database. Filtering and faceting of data are supported.
Data endpoints	A comprehensive range of data endpoints are offered, including SPSS, Stata, Statistica, SAS, and Dbase. Users can also download data in Excel, PDF or CSV. These are also available via the API.
Visualisation	While not striking, Nesstar offers a comprehensive range of visualisation and analytical functions which can be applied to datacubes. Charts can be saved and shared, or embedded into a CMS. Beyond charting, Nesstar offers users the option of adding in new calculations into tables as well as performing correlation analysis.
UX & S/W extensibility	With a comprehensive API and many open source components, Nesstar would appear suitable for some customisation, however little is reflected in the various independent sites running the software. Most appear identical.

Observations

Nesstar was originally designed to support microdata publication and does not meet all the criteria required for an open data portal, with a need for individual dataset licensing, improved search, and metadata faceting. Nesstar does provide comprehensive data support, including individual files and data-series, as well as integrated datacubes and is well suited to the immediate requirements of NSOs. Few other platforms are as well integrated.

5.7 OpenDataSoft

URL	http://www.opendatasoft.com/
S/W Licence	Proprietary
Language	
SaaS	http://www.opendatasoft.com/
Demo	http://public.opendatasoft.com/
Examples	opendata.brussels.be
	opendata.paris.fr
	data.sncf.com

Overview

OpenDataSoft offers a comprehensive suite of open data and visualisation tools. Their search functionality is straightforward with faceting / filtering and well-structured results listings which include icons defining the alternative forms for the data (such as table, map, charts or export).

Open data suitability

Descriptive metadata	OpenDataSoft supports DCAT, and INSPIRE for geospatial data. You are also able to create custom metadata templates.
Machine-readable	A wide range of data-types are readable by the software, including Shapefiles, OSM, KML, WFS, ESRI, GTFS, as well as the more traditional XLS, CSV and XML types. There is also the potential to link to alternative data sources, such as web forms, other databases, and APIs.
Anonymous access	Users are able to access the site anonymously, including downloading data and creating visualisations. An API is similarly available for anonymous querying and downloading, although a key will be required for modification of data.
Data licenses	All licenses are clearly referenced, including in search results.
Data attribution	Similarly to licenses, attribution is clearly referenced, including in search results.
Search	Natural language search is available, including filtering by a wide range of metadata and data-types. The API similarly permits faceting during search.
Open API	The API is open and includes an interactive online dashboard, plus clear documentation, for testing and working with the API. The API permits HTTP/HTTPS/BasicAuth and presents data in JSON/P, CSV, RDF, as well as GeoJSON/P
Static URI	Static URIs are plentiful. Anonymous users can create custom visualisations and share the URI for this. URIs are generated for every different view ensuring easy sharing and referencing.
Harvesting	OpenDataSoft has a range of services for importing data from a wide range of services, including setting processes for removing personal data, performing calculations based on formulae. Data collection can be via remote locations or web services.
Federating	The software is provided as SaaS but each site is currently presented independently. No federation takes place at this time, but the API and metadata traversal means that this should be possible in future.
Public Documentation	The public documentation on use of the API is fairly good (http://public.opendatasoft.com/api/doc/), but there is little on the software itself, from user to administration. Most guidance is provided via videos on the main site. http://www.opendatasoft.com/ressources/
Standards-based	With support for RDF and the leading open data metadata

standards, OpenDataSoft is aligned with industry standards.

NSO suitability

Structural metadata	OpenDataSoft currently provides no support for structural metadata.
OLAP hypercube	Currently, there is no hypercube support, however they are in development of an OLAP based on Microsoft's **MDX**[34] standard.
Data endpoints	Beyond the standard endpoints of CSV, XML and JSON, and geospatial formats like KML, WFS, GTFS, they have also developed an OData endpoint as well.
Visualisation	There is very good visualisation support, including the ability to embed visualisations in other web services. Images cannot be exported as PDF or image files. Chart-types include line, spline, column, area, bar and pie charts. OpenDataSoft has also developed a comprehensive geospatial visualisation platform called Cartograph which permits multiple geospatial data to be presented simultaneously. The map can then be shared or embedded.
UX & S/W extensibility	They are a SaaS vendor and, as with all such vendors, deployment is straightforward. Their template interface and customisation options appear fairly good and the various client sites are quite different from each other. However, full customisation is limited to the OpenDataSoft team.

Observations

OpenDataSoft is one of the more sophisticated open data platforms and well designed to serve that need. While it is not yet entirely suitable for NSO requirements, they are currently preparing an OLAP extension which will also lead to support for structural metadata.

5.8 PC-Axis and PX-Web

URL	http://www.scb.se/pc-axis
S/W Licence	Proprietary
Language	Unknown
SaaS	No
Demo	statistikdatabasen.scb.se
Examples	www.bfs.admin.ch
	www.cso.ie
	www.stats.govt.nz

Overview

The PC-Axis family consists of a number of programs for the Windows and Internet environment used to present statistical information. It is mostly used by the statistical offices in different countries to let their users retrieve statistics.

PC-Axis is a software family with several programs all aimed at facilitating quick and easy dissemination of statistics. PX-Web is the online data publishing and presentation component of PC-Axis.

Open data suitability

Descriptive metadata	In PC-Axis, such metadata about data objects is referred to as the quality data and this is supported as a set of additional views. This supports descriptions and definitions for the data.
Machine-readable	PC-Axis is designed to inform a hypercube and the system only accepts PC-Axis format files in order to import the necessary data. General machine-readable files are not supported.
Anonymous access	Anonymous user access to data via the PX-Web browser is supported, as well as administrative permissions for managing the data itself.
Data licenses	Licensing is presented at site level since the assumption is that all data are released from the same source. This is not necessarily always the case and does limit multi-organisation, multi-licensing releases.
Data attribution	Once data are extracted from the hypercube, metadata are provided with clear attribution and contact details for the series selected.
Search	The mechanism for searching datasets is an interactive statistical browser in which dates, data series and geographical range are selected prior to data being presented. It is equivalent to performing a SQL data lookup. This means that users require specialist knowledge about the classification of statistical data in order to find data of interest. PX-Web permits individual pages of text, with tables and special views, to be created but this does not permit searching to find data of interest.
Open API	PX-Web has begun offering an API, although this is not offered across all deployments. Statistics Sweden's API and documentation is available here: http://www.scb.se/en_/About-us/Open-data-API/API-for-the-Statistical-Database-/. Data output is for a range of formats, including: PX (PC-Axis), CSV, JSON, XLSX, JSON-STAT and SDMX.
Static URI	Static URIs to specific data series are not possible.
Harvesting	Data files and resources which are in PC-Axis format can be harvested automatically from remote folders.
Federating	The service is not designed for federation.
Documentation	Documentation on the public PC-Axis file structure is readily

	available via PDF documents from the main PC-Axis website although there is no public developers documentation site. Licensees of PX-Web are permitted access to further documentation and are able to access the source code.
Standards-based	PC-Axis has become a widely-used standard for NSOs, and PX-Web similarly supports SDMX with some support for DDI as well. The software itself is proprietary.

NSO suitability

Structural metadata	There is a range of support for data structure metadata, including PC-Axis for internal management, SDMX for data exchange, and DDI for some data formats.
OLAP hypercube	PC-Axis has comprehensive hypercube support. The hypercube supports full filtering and faceting of dataseries, and similar functionality is accessible via the API.
Data endpoints	The web interface and API offer a range of machine-readable file-formats as endpoints, including PX (PC-Axis), CSV, JSON, XLSX, JSON-STAT and SDMX.
Visualisation	PX-Web does not provide full business intelligence functionality but does offer a series of static options for data visualisation, including tables, line, bar and pie charts, and some geographic representation. This is not a fully interactive visual package but does provide quick and clear functionality. Charts are not shareable or persistent for end-users.
UX & S/W extensibility	Licensees of the platform have access to the source code and are able to customise the user-interface. The API permits software extensibility while licensees are similarly able to extend the software platform as required.

Observations

PC-Axis meets the needs of NSOs for data publication, hypercubes and metadata, but requires enhancements in servicing open data needs. It can only store PC-Axis-compliant data and import these into hypercubes, but not generic data files nor metadata to support these. There is also no URI for data address persistence.

PC-Axis requires a degree of expert knowledge to use and this limits data discovery for lay users.

5.9 Prognoz

URL	http://www.prognoz.com/
S/W Licence	Proprietary
Language	Unknown
SaaS	http://www.prognoz.com/
Demo	http://dataportal.prognoz.com/
Examples	nigeria.prognoz.com
	indicators.statistics.gov.rw
	dataportal.afdb.org

Overview

Prognoz is a business intelligence platform which supports the development of software solutions on the desktop, web, and mobile devices for visualisation and OLAP, reporting, and modelling and forecasting of business processes.

The Prognoz Platform provides collaboration with portal solutions such as MS SharePoint, SAP Netweaver, IBM WebSphere, and GIS services such as Google Maps, Microsoft Bing, OpenStreetMap, and Yandex Maps.

Open data suitability

Descriptive metadata	Prognoz uses their own metadata structure but it is customisable so could be set up to mimic standard approaches.
Machine-readable	An ETL module allows data import from databases such as Oracle, Microsoft SQL Server, IBM DB2, and a variety of different file-types XML, EDIFACT, DBF, TXT, and XLS/X. At this stage there doesn't appear to be much support for geospatial data.
Anonymous access	Data access can be strictly controlled but public datasets are available without a requirement for a login.
Data licenses	There is no clear metadata specifying the license for reuse of the underlying data.
Data attribution	A generic link to the data source is available.
Search	Search can be implemented in a number of different ways, including full text-search plus filtering based on metadata, or implemented only as a branching tree system for data traversal. Search results are sometimes presented with a degree of interactivity where it is not always clear how to download or access the data. Mostly, however, results take you through to a data interaction page.
Open API	There is no public API for accessing the data or building new functionality.
Static URI	There are no static URIs to any individual pages within the portals; none of search results, data pages, or custom views generate a shareable URI.
Harvesting	Prognoz has a visual ETL task manager which permits creation of processes for transformation and loading of data. These can be set to run regularly as data changes. Data can also be set to be loaded directly into datacubes.
Federating	Technically, since Prognoz has a sophisticated harvesting system, it can harvest from other Prognoz instances – only importing the metadata – and acting as a federated system.
Public Documentation	Public documentation is minimal. Given there is no public API, this is expected.
Standards-based	Prognoz integrates well with Microsoft Office and related services but does not comply with either of open data or

	statistical data publication standards. Prognoz is mainly aimed at proprietary commercial business intelligence requirements rather than interoperability or standards.

NSO suitability

Structural metadata	Similarly to with the descriptive metadata, Prognoz has its own approach to managing structural metadata.
OLAP hypercube	Full support for hypercubes are available, including the ability to facet and produce custom slices. As a business intelligence tool, cubes can also be subjected to numerous transformations including analysis and forecasting, validation and so on.
Data endpoints	Data selections can be downloaded as any of XLS, XLSX, PDF, RTF, HTML, MHT, PPTX, ODS, EMF, and PPReport.
Visualisation	Prognoz is a comprehensive business intelligence platform with the ability to conduct analysis, including modelling and forecasting, on the data, as well as producing complex visualisations and dashboards. While there doesn't appear to be support for shapefiles or other coordinate data, place-names are recognised and plotted on maps. Searchable dashboards permit visual data exploration, and a wide range of endpoints allow for the charts and graphics to be downloaded for presentation elsewhere.
UX & S/W extensibility	Prognoz comes with its own software development toolkit which is compatible with .NET. This permits developing macros for data management, forms for online data capture, as well as creating custom visualisations and charts. External libraries for data representation using technologies of COM, ActiveX, Flash, .NET, and ASP.NET can also be used. Note that none of the development tools or documentation are public or standard and so third-party developers are unlikely to have experience with the software.

Observations

Prognoz is not suitable as an open data portal, offering few of the requirements for such data publication. It also does not appear to be entirely suitable for NSOs since it does not support standard metadata requirements.

Prognoz is clearly aimed at the bespoke needs of the corporate environment and has little support for the standards taken for granted in more collaborative industries. While it has been used in data publication it appears to be an unusual choice given the amount of data transformation required for interoperability with other statistical platforms.

5.10 Semantic MediaWiki

URL	http://semantic-mediawiki.org/
S/W Licence	GNU General Public License
Language	PHP
SaaS	http://www.referata.com/
Demo	http://semantic-mediawiki.org/
Examples	openei.org
	floridalegalwiki.com
	www.skybrary.aero

Overview

Semantic MediaWiki is an extension of MediaWiki – the wiki application best known for powering Wikipedia – that helps to search, organise, tag, browse, evaluate, and share the wiki's content. While traditional wikis contain only text, SMW adds semantic annotations that allow a wiki to function as a collaborative database.

Semantic MediaWiki is the only data publication platform evaluated in this report offering version control.

Open data suitability

Descriptive metadata	Semantic MediaWiki is an RDF implementation with templates to structure the metadata linked to imported data.
Machine-readable	Importing of data is performed via XML or CSV only, with additional extensions permitting JSON as well. The expectation, though, is that data is read using the CSV format for inline queries. There is also recognition of coordinate data for plotting on maps.
Anonymous access	Users may remain anonymous but there is a full authentication service as well for data and interaction management.
Data licenses	Metadata templates present data licenses on each page.
Data attribution	Similar to licenses, each dataset is presented with its source.
Search	Free text search is supported, although search results are limited to title and an excerpt from the description. Filtering is limited to preset metadata which does not guide the user to refining the results to any great degree.
Open API	The platform is queryable via a SPARQL interface and is able to return JSON data serialisation. Note, though, that the API only queries the database. Extending the software is done via independent modules that must be plugged into the software itself.
Static URI	Static links are available for all data and views.
Harvesting	There is limited support for automated importing of data.
Federating	As with harvesting, federation of data sites is limited.
Public Documentation	Semantic MediaWiki has been in continuous development since 2005 and has a large and enthusiastic developer community. As with all popular open source projects, documentation is comprehensive and widely available.
Standards-based	The software has a passionate community and numerous research projects and extensions have aimed to ensure that software is entirely standards compliant. Many of their researchers are statisticians and have aimed to develop features of use to NSOs.

NSO suitability

Structural metadata	Semantic MediaWiki does not currently provide support for structural metadata.

OLAP hypercube	Papers offering proof-of-concept for OLAP support for Semantic MediaWiki have been published but formal implementations have not yet been completed. The likelihood is that support would be via the RDF Data Cube vocabulary.
Data endpoints	The system provides output as XML, JSON and via SPARQL.
Visualisation	Visualisation is extremely limited and mostly to tabular formats, but extensions can be developed for a variety of open source libraries.
UX & S/W extensibility	MediaWiki is a platform in its own right and a vast number of software extensions have been written to enhance it. Similarly, the active developer community has written up comprehensive documentation which is available to support any custom extension or UX work which may be required.

Observations

Semantic MediaWiki is suitable as an open data publication service. While there are initiatives underway to incorporate NSO requirements, it is does not currently meet those needs.

Semantic MediaWiki is still fitted to MediaWiki which means a fairly rigid template style and that the service is mostly about text. The WikiData project is aimed mainly at data but currently for internal MediaWiki use. The likelihood is that these two projects will start to merge.

Despite an extremely large developer community, and numerous working NSO statisticians amongst the developers, there are few NSO portals built on this platform.

5.11 Socrata

URL	http://www.socrata.com/
S/W Licence	Mixed proprietary and OS
Language	Scala, Javascript, Ruby
SaaS	socrata.com/products/open-data-portal/
Demo	nycopendata.socrata.com
Examples	data.undp.org
	data.cityofchicago.org
	opendata.go.ke

Overview

Socrata's Open Data Portal SaaS provides one of the more comprehensive open data services, with a range of extensions for dashboards, live reports and the ability to manipulate and update existing data live in the portal.

Their commitment to SaaS means you can deploy a new site and be serving data in a day. Beyond open data, they offer business intelligence and visualisation functionality permitting data visualisation, analysis and sharing via social media.

Open data suitability

Descriptive metadata	Socrata uses RDF metadata to describe the datasets, presented in Dublin Core and DCAT, as well as custom metadata fields.
Machine-readable	Socrata is able to read and produce the following data types: CSV, JSON, PDF, RDF, RSS, XLS, XLSX, XML, OData, Shapefile, KMZ, and KML.
Anonymous access	Users do not need to create accounts in order to access data, but they will need to do so in order to access higher services. Such include following datasets, commenting and saving visualisations that they produce from the data.
Data licenses	Each dataset is individually licenced and is clearly labelled.
Data attribution	Socrata does present clear attribution when such data exists.
Search	Searching is fast and user-friendly, with the ability to filter by view types, as well as categories and topics. Information returned in the search offers not only the description of each dataset, but also an abbreviated view of the first three matching rows of data, permitting rapid assessment of results. Datasets can be filtered and facetted via the web interface as well as the API.
Open API	Socrata produces a wide range of endpoints via their API, including REST JSON, CSV and RDF-XML. Their documentation is comprehensive http://dev.socrata.com/docs/endpoints.html and permits developers to experiment with the system live. Authentication is required for users wishing to push data to Socrata's servers.
Static URI	Each dataset and view gets its own URI as well as a generated short URI to facilitate sharing via social media.
Harvesting	The API permits development of automated processes for uploading fast-changing datasets or importing existing resources. Socrata provides a dashboard for managing such processes as well. Socrata supports The White House's /data.JSON URL extension specification.
Federating	Since all Socrata sites run on a single server, federating and sharing resources/datasets between Socrata sites is straightforward. Socrata has developed additional extensions to import metadata from alternative open data portals, such as CKAN.

Documentation	Possibly Socrata's greatest strength is the well-developed and presented developer portal including numerous libraries for working with software as diverse as the R statistical platform, Scala, Ruby and Java, amongst others. Their developer portal is available at http://dev.socrata.com/.
Standards-based	Socrata has adopted a mixed licensing approach with their core architecture for their centralised, scalable systems being proprietary and the various tools available via their API available open source. Most of the individual software components for Socrata are open source, available on their Github repository (Socrata Open Data Server Community Edition), and data storage and presentation complies with the emerging open data standards, such as RDF.

NSO suitability

Structural metadata	Socrata does not provide standardised metadata for dataset structure or format – though publishers can set custom metadata fields.
OLAP hypercube	Socrata does not support presentation of data as a hypercube. The software approach does facilitate eventual hypercube support as all machine-readable data are imported into a database, permitting column-alignment. Similarly, metadata can be edited and this could be enhanced to support NSO requirements.
Data endpoints	Socrata provides a range of endpoints, including OData, JSON and XML. They do not provide any of the proprietary formats, such as Stata or SPSS.
Visualisation	While Socrata is not yet a full business intelligence service, the range of visualisations is extensive. Chart creation capabilities includes various chart types such as Area, Bar, Column, Donut, Line, Pie, Time Line, Tree Map and Heat Map. Geospatial support and visualisations include location data, or GIS files such as Esri shapefiles, KML/KMZ files, using either Google Maps, Bing Maps or ESRI. A range of additional interfaces make live report generation and charting straightforward even for the layman.
UX & S/W extensibility	Socrata is designed to be easy to deploy and is managed as SaaS. This reduces complexity in management but also limits the degree to which sites can be customised. Landing pages can certainly be bespoke, but interactivity in search and visualisation remains quite consistently defined. Sites can have a degree of colour and branding changes, but the overall look-and-feel remains similar. Given the breadth of interactivity possible via the API, extending Socrata is straightforward. A library of existing extensions, released under various open source licenses (including the liberal MIT license) are available on their Github repository at https://github.com/socrata.

Observations

Socrata is a good choice for open data sites but will require development to support hypercubes and dataset-level structured metadata in order to support the complete requirements for an integrated NSO portal.

5.12 Swirrl

URL	http://www.swirrl.com/
S/W Licence	Mix proprietary & open source
Language	Rails
SaaS	swirrl.com/publishmydata
Demo	
Examples	opendatacommunities.org
	opendatascotland.org
	linkeddata.hants.gov.uk

Overview

Swirrl's PublishMyData platform is probably the purest linked data service available for publishing open data. RDF with SPARQL are still sufficiently novel that many data publishers do not necessarily think about data architecture when deciding on their vendor. Fully-realised RDF offers the most future-proof mechanism for data publishing and is worth considering.

Swirrl also offers a mixed proprietary and open source set of licenses. Their hosted environment and configuration is proprietary while they offer numerous Ruby-based libraries for developers using the GNU Affero General Public License or MIT license.

Open data suitability

Descriptive metadata	Swirrl offers RDF as the mechanism for metadata. This is extensible and underlies common metadata formats like DCAT or Dublin Core.
Machine-readable	Standard machine-readable formats like CSV and XLS/X (Excel) are supported. If the data are machine-readable then Swirrl will serve it, however, there is limited recognition of datatypes (e.g. geospatial and similar).
Anonymous access	Users have anonymous access and there are a wealth of controls for authentication management.
Data licenses	Individual data are clearly licensed.
Data attribution	Attribution for every dataset is implemented.
Search	Swirrl does not have a user-interface for search although the data can be traversed and searched via the API. At present, data are simply listed via an interface. A text search facility is scheduled for the next version of the software.
Open API	The API is a SPARQL implementation offering a standards-based interface for interaction and application development.
Static URI	All data and endpoints are offered via static URIs making sharing and referencing straightforward.
Harvesting	The API offers full interaction and data manipulation, and the SaaS platform offers dashboards for uploading and importing files. API keys are required for authenticated actions.
Federating	There is currently limited support for federating from non-RDF sites, but the API does permit importing metadata from other RDF-compatible services.
Public Documentation	There is fairly good documentation on using the API (http://opendatacommunities.org/docs) although there isn't very much public information on the interface for the SaaS or for the open source version of the software.
Standards-based	Swirrl is the purest implementation of RDF for open data currently available. It adheres closely to W3C standards.

NSO suitability

Structural metadata	PublishMyData incorporates a number of tools for data using the W3C Data Cube Vocabulary and Swirrl is involved in a research project to extend RDF Data Cube support[35].

OLAP hypercube	The system offers tools for transforming CSV files and Excel spreadsheets to RDF Data Cube datasets (without the need for programming), managing a collection of concept schemes, selecting URIs from external reference data, and quality checks that make sure any generated RDF meets the required standards.
Data endpoints	The API offers JSON and SPARQL as endpoints.
Visualisation	There is no native visualisation system but the API permits integration with various other visualisation systems. Basic native visualisation is scheduled for the next release.
UX & S/W extensibility	The community edition of Swirrl permits full customisation (http://github.com/swirrl/publish_my_data) while the online platform SaaS can also be customised or integrated into other systems.

Observations

Swirrl is designed for open data publication and meets the requirements of these services. It does not yet, however, meet all the requirements for NSO portals.

They are committed to linked data with a focus on technical users of statistical data, and offer RDF Data Cube support. This offers a comprehensive data publication and management service.

Recognise, though, that Swirrl is currently aimed at technical users and developers, but is working on enhanced features for less technical users.

6 Conclusions and recommendations

The software platforms considered in this report are not a comprehensive list of those available to NSOs and open data publishers. Even so, many of them come close to meeting all the criteria for both requirements.

Open data software has led to a number of different approaches to resolving requirements for data publication, engagement and reuse. Many of these do not promote interoperability between platforms and new problems are being raised.

Here are a list of attributes where enhancement would improve the overall utility to NSOs and the open data community.

6.1 Improve technical documentation

Too little of the available documentation for developing custom components or using the software APIs offers much support to developers. Worse, many software services release no documentation at all to the public.

Open data is not only about the release of content but also of the mechanisms to engage with that content.

This is something where the ostensible open data solutions do reasonably well, although some could be improved. However, releasing documentation in poorly-updated PDFs is unacceptable and of limited use.

If software vendors persist in developing proprietary and incompatible APIs then it is essential that these APIs be sufficiently well-documented that they can be of use.

6.2 Ensure public APIs and endpoints are interoperable

In addition to providing complete documentation, it is recommended that vendors adopt consistent and interoperable APIs.

At present, a developer wishing to integrate two unrelated software platforms has to manually craft a solution which has to read two different APIs.

If vendors are able to agree on a common API – or are able to connect to a common standard – then harvesting from different systems as well as developing applications that integrate a number of unrelated platforms all become extremely straightforward.

Adopting RDF is one mechanism which will permit eventual harmonisation. However, this also means agreeing to metadata standards and terms which permit different software platforms to communicate directly. Going from human-readable to machine-readable inevitably requires compromise but it is essential.

Critically, it isn't helpful if RDF is used for data architecture (i.e. to find out what is in the database of datasets in the platform) but there is still no standardised way to traverse that architecture and arrive at a data endpoint.

The likelihood is that NSOs will not manage all the data in a particular country and that different data publishers will choose different software platforms. It is essential that harvesting and federation adopt common standards for interoperability.

This will permit services like IPUMS to harvest census data directly without have to gather direct copies and restructure the data. It will also reduce multiple software systems serving the overlapping needs of funders and stakeholders.

Such interoperability will permit NSOs to adopt a staged approach to updating and integrating their systems. Retrofitting DDS to existing systems will be more straightforward if those systems are able to pull data for visualisations and dashboards directly from an RDF endpoint.

This does not imply that SDMX, DDI or OData are necessary solutions, but they may provide interim points towards SPARQL and RDF data cubes.

Common endpoints and structural metadata, at least, permit third-party software to interpret data, when it is eventually discovered.

6.3 Presentation of metadata and URIs must conform to W3C standards

It is critical that data be released in machine-readable format but it is just as important that search and data discovery generate unique and persistent URIs.

Metadata associated with datasets needs also to be easily discoverable and presented with the data it applies to.

Many software platforms fail to present descriptive metadata which aids discovery and reuse. It needs to be clear what the licensing and reuse policies are, what the data are about, and who is responsible for it.

Similarly, data discovery is time-consuming and discrete URIs for each step of the process permits sharing and saving of these states. A user who spends half-an-hour finding data only to be unable to simply send a link to colleague (or bookmark it for later use) is less likely to engage with the data at all.

This is simple compliance with W3C standards for web acceptable applications and is not specific to open data. Best practice, in this case, is simply good internet manners.

6.4 Natural language search and metadata faceting should be standard

Many of the solutions chosen for data publication favour visualisation and presentation over discovery and reuse.

From a budget allocation perspective, plumbing usually takes second-place to more attractive considerations. Sadly, this makes data discovery awkward and reduces the potential for data reuse.

Google, Bing and Yahoo are the most common search experience for most users. Expert systems, which data software are, can enhance the search experience through permitting metadata to be used as additional context-sensitive filters.

Data discovery should not be limited to professionals who are already familiar with the data. Free text search along with metadata faceting speeds up data discovery and improves system performance.

6.5 Structural metadata and hypercube support are core NSO requirements

Statistical data are used to develop research insight. Individual tables become more useful when they can be aggregated together and sliced into numerous views for analysis.

Obviously if the data do not conform to a structural metadata standard such as SDMX, PC-Axis or DDI, then offering support for hypercubes is not going to be useful. However, in the case of NSOs, supporting commonly-used structural metadata is common.

With the release of the W3C's RDF Data Cube Vocabulary, it is also becoming less necessary to build a complete implementation from scratch.

For some open data software services this will prove an extremely difficult enhancement but offering support for both descriptive and structural metadata, as well as hypercubes, is all part of ensuring the data are as widely available and as useful as is possible.

6.6 Dashboards and visualisations are necessary for user engagement

User-engagement in the face of vast and complex data is more likely where that data are presented in a tactile and engaging way. The growth of data journalism has depended on the growing availability of public data resources and their ability to present visually exciting views on that data.

Numerous platforms already offer data visualisation, dashboard development and some level of business intelligence. For others, where such systems require major investment, compliance with standards and interoperability will permit integration with existing software.

Online visualisation systems weren't evaluated for this report, but there are a vast number, ranging from simple solutions like Datawrapper.de to sophisticated business intelligence systems like Tableau.

If software cannot offer everything, then it should offer integration.

6.7 Develop data engagement tools for improving data-quality and reuse

Visualisations permit users to interact with the data but not necessarily support the enhancement of data, or its reuse.

There are a number of approaches which promote data quality and reuse:

- **Use the data you publish**: data publishers should have a mechanism in the software to derive visualisations and analysis from the data, and save or link these to published datasets;
- **Showcase user applications:** a mechanism for users to share their applications, research or content developed, with a workflow for site administrators to evaluate and present such content;
- **Register issues:** instead of comments which require moderation, offer users the opportunity to raise issues with data quality for each dataset; provide a tracker for response and a dashboard showing issue responses;
- **Data requests:** offer a direct mechanism for users to request datasets which may not yet be available; Data.gov.uk offers known – but unavailable – data in search results but flagged as requiring a formal request;

7 Acknowledgements and Research Methodology

This technical research assessment was commissioned and supported by the World Bank. The report was researched and written by Gavin Chait of Whythawk[36], open data software consultants.

Both primary interviews and secondary research of existing literature contributed to the content of this report.

The complete list of people interviewed as part of the study (in alphabetical order of organisation or software) are:

- Adam McGreggor and Adrià Mercader of the CKAN team at Open Knowledge;
- Andrew Hoppin of the DKAN team at NuCivic;
- Matthew Welch and Olivier Dupriez on behalf of the International Household Survey Network (IHSN);
- Robert McCaa at IPUMS;
- Diego May at Junar;
- Jean-Marc Lazard at OpenDataSoft;
- Rajiv Ranjan at Rwanda's National Statistics Office;
- Ben McInnis, Joe Pringle, Jessica Carsten and Jeff Kaplan at Socrata;
- Bill Joyce at Statistics Canada (Canada NSO);
- Lars Knudsen at Statistics Denmark (Denmark NSO);
- Bill Roberts at Swirrl;
- Marian Brady, Oliver Fischer and Jeffrey Sisson of the US Census Office;
- Tim Harris, Tim Herzog and Thomas Danielewitz at the World Bank;

All secondary sources are cited in the text and available in the References section.

Analysis presented in this report is derived over the course of the interviews conducted during the primary research phase and offer insight software and design choices and priorities. It cannot be considered a statistically relevant sample, but does provide a sense-check as to how different organisations and entities across the open data industry respond to constraints and opportunities.

8 Glossary

Application Programming Interface (API):	Specifies how some software components should interact with each other. Used to ease the work of programming graphical user interface components, to allow integration of new features into existing applications, or to share data between otherwise distinct applications. Presented as a library that includes specifications for routines, data structures, object classes, and variables. In some other cases, notably for SOAP and REST services, an API comes as just a specification of remote calls exposed to the API consumers.
Business Intelligence Systems (BIS):	A platform for engaging with structured quantitative data to produce custom slices of that data, charts, tables and geospatial representations.
Content Management Systems (CMS):	Permit publishing, editing and modifying of qualitative content, as well as providing mechanisms to manage workflows and individual users in a collaborative environment.
Data Discovery Systems (DDS):	Similar to CMS but provide mechanisms to manage the semi-structured quantitative and qualitative data in documents and spreadsheets and offer methods for data publication, discovery and reuse.
Datastore:	A data repository of a set of integrated objects modelled using classes defined in database schemas. A datastore includes not only data repositories like databases, but more generally includes also flat files that can store data.
Descriptive metadata:	Corresponds to external metadata and typically used for discovery and identification, as information used to search and locate an object such as title, author, subjects, keywords, publisher.
Extract, Transform and Load (ETL):	A process in database usage and especially in data warehousing that: extracts data from outside sources; transforms it to fit operational needs, which can include quality levels; loads it into the end target (database, more specifically, operational data store, data mart, or data warehouse).
Faceting:	Also called faceted search, faceted navigation or faceted browsing, is a technique for accessing information organised according to a faceted classification system, allowing users to explore a collection of information by applying multiple filters. Each information element is classified along multiple explicit dimensions, enabling classifications to be accessed and ordered in multiple ways rather than in a single, pre-determined, taxonomic order.
Federation:	A meta-database management system, which transparently maps multiple autonomous database systems into a single federated database. The constituent databases are interconnected via a computer network and may be geographically decentralised. Since the constituent database systems remain autonomous, a federated database system is a contrastable alternative to the (sometimes daunting) task of merging several disparate databases. A federated database, or virtual database, is a composite of all constituent databases in a federated database system. There is no actual data integration in the constituent disparate databases as a result of data federation.
Filestore:	A collection of binary data stored as individual files and referenced in a database management system.
Generalised data:	Aggregations derived from microdata; for example, the total number of people of a particular education category.
Harvesting:	An automated and autonomous mechanism for ETL of known data from

	known web addressable locations into a single database or datastore.
Linked Data:	A method of publishing structured data so that it can be interlinked. It builds upon standard web technologies such as HTTP, RDF and URIs extending them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried.
Microdata:	Information at the level of individual respondents, households and businesses, typically through surveys; for example, a national census may collect age, address, education, employment status, etc. from individuals.
Online Analytical Processing (OLAP) hypercube or cube:	An approach to enable users to analyse multidimensional data interactively from multiple perspectives. OLAP consists of three basic analytical operations: consolidation (roll-up), drill-down, and slicing and dicing. Consolidation involves the aggregation of data that can be accumulated and computed in one or more dimensions. Drill-down is a technique that allows users to navigate through the details. Slicing and dicing is a feature whereby users can take out (slicing) a specific set of data of the OLAP cube and view (dicing) the slices from different viewpoints.
Open Source Software (OS):	The source-code is available in an online and public repository under a liberal reuse license (such as the General Public License and its affiliates); sometimes known as Free and Open Source Software (FOSS), not all open source software is free, and not all free software is open source.
Representational State Transfer (REST):	An architectural style consisting of a coordinated set of architectural constraints applied to components, connectors, and data elements, within a distributed hypermedia system. REST ignores the details of component implementation and protocol syntax in order to focus on the roles of components, the constraints upon their interaction with other components, and their interpretation of significant data elements.
Resource Description Framework (RDF):	General method for conceptual description or modelling of information that is implemented in web resources, using a variety of syntax notations and data serialisation formats.
Semantic interoperability:	The ability for computer systems to exchange data unambiguously.
Software-as-a-Service (SaaS):	Software is available online on a centralised hosted server via a subscription service instead of as a deployable software system with a single, static price; upgrades, bug fixes and patches are consistently and regularly applied; custom extensions of functionality can be achieved via an API but customisation of the user interface is more limited.
Structural metadata:	Corresponds to internal metadata about the structure of database objects such as tables, columns, keys and indexes.
Uniform Resource Identifiers (URI):	A string of characters used to identify a name of a resource. Such identification enables interaction with representations of the resource over a network, typically the World Wide Web, using specific protocols. Schemes specifying a concrete syntax and associated protocols define each URI. URIs can consist of both namespaces and locators at the same time.
Uniform Resource Locators (URL):	Also known as web address, particularly when used with HTTP, is a specific character string that constitutes a reference to a resource. In most web browsers, the URL of a web page is displayed on top inside an address bar. A URL is a type of URI.

9 References

[1] Open data challenges and opportunities for national statistical offices. Washington, DC: World Bank Group. World Bank. 2014. http://documents.worldbank.org/curated/en/2014/07/19791395/open-data-challenges-opportunities-national-statistical-offices-open-data-challenges-opportunities-national-statistical-offices
[2] Ibid
[3] Technology Options for Open Government Data Platforms – Timothy Herzog, World Bank, 2014-01-31
[4] https://international.ipums.org/international/
[5] http://www.ihsn.org/home/software/nada
[6] http://www.w3.org/RDF/
[7] Anwar and Hunt BMC Bioinformatics 2009 10(Suppl 10):S3 doi:10.1186/1471-2105-10-S10-S3
[8] https://www.gnu.org/licenses/licenses.html
[9] http://www.w3.org/TR/vocab-dcat/
[10] http://www.ddialliance.org/
[11] http://dublincore.org/documents/dces/
[12] http://inspire.ec.europa.eu/index.cfm
[13] ISO 19115-1:2014, Geographic information -- Metadata http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798
[14] Linked Data – Tim Berners-Lee, 2006-07-27 http://www.w3.org/DesignIssues/LinkedData.html
[15] New paywall costs the Times 66% of its internet readership, http://www.theguardian.com/media/2010/jul/18/times-paywall-readership
[16] Showing you this map of aggregated bullfrog occurrences would be illegal, http://peterdesmet.com/posts/illegal-bullfrogs.html
[17] http://creativecommons.org/licenses/
[18] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/
[19] http://opendatacommons.org/licenses/odbl/1.0/
[20] http://creativecommons.org/licenses/by/3.0/igo/
[21] The Semantic Web - Tim Berners-Lee, James Hendler and Ora Lassila, Scientific American, 2001-05, http://www.scientificamerican.com/article/the-semantic-web/ [Subscription, annoyingly]
[22] That's just semantics - Gareth McGuinness, International Monetary Fund, 2009-10, https://app.box.com/shared/9zms2mvsio
[23] http://www.scb.se/sv_/PC-Axis/About-PC-Axis/
[24] http://sdmx.org/
[25] https://dvcs.w3.org/hg/gld/raw-file/default/data-cube/index.html
[26] http://www.w3.org/TR/owl-features/
[27] http://data.gov.uk/resources/coins
[28] http://www.json.org/
[29] http://www.odata.org/
[30] http://stats.oecd.org/OpenDataAPI/Index.htm
[31] http://www.w3.org/XML/
[32] http://www.w3.org/TR/sparql11-overview/
[33] http://www.openarchives.org/OAI/openarchivesprotocol.html
[34] http://en.wikipedia.org/wiki/Multidimensional_Expressions
[35] http://blog.swirrl.com/articles/open-cube/
[36] http://www.whythawk.com/