The role of Information Asset Registers

1. Background

The UK Government's aim is to unlock the benefits of public sector information (PSI) re-use, opening up its data for both economic and social gain. The Office of Public Sector Information (OPSI) is responsible for the UK Government's information policy including the management of Crown copyright. OPSI aims to identify strategies and approaches that enable serendipitous re-use. Public sector information is a major factor driving innovation on the web. Web-enabled PSI supports new forms of social production of information and knowledge. The availability of public sector information is essential to supporting the type of cumulative innovation required in a knowledge economy.

Governments have much information which people want to access – either to find and use or, more importantly, to re-use. Increasingly, the most useful services are those which combine data from different sources, mixing public, private and user created content. This process of "data mashing" is happening on the web, using web technologies, standards and approaches.

2. The Power of Information

In the UK the Steinberg / Mayo Power of Information Review published in 2007 makes a powerful case for government engaging with online communities and releasing more of its data. It is an agenda which has been embraced enthusiastically. Ministers such as Tom Watson, responsible for "Transformational Government" at the Cabinet Office, and Michael Wills, the Minister for PSI, talk comfortably about information as infrastructure. A new Power of Information Taskforce is pushing forward initiatives such as the *Show Us a Better Way* competition. In return for a £20,000 prize, the competition asks "*Do you think that better use of public information could improve health, education, justice or society at large?"* and invites people to submit their ideas online.

The web is underpinning this whole process. It provides a way for re-users to source the ingredients for new types of service. Web standards, such as HTTP, XML, RDF and URIs, provide the essential tools which allow the data to be combined. The web also provides the presentation layer, the place where people come to access new sources of information and new services.

Information is not only being re-used in increasingly sophisticated combinations, it is also being linked together. The process of linking web pages together is as old as the web. Understanding the network of links and link texts is an essential part of Google's page rank algorithm, for example. The emerging web of linked data is linking information at the fragment level, rather than the web page. This introduces a form of reliance between information providers. As a provenance source of reliable, official information and data, governments potentially have an important role as part of this emerging web of linked data.

A key point recognition is that a government website is not always the most effective place to provide information to the end user. Better that the information is where the users are, which means its re-use by others on the web. For example, food hygiene inspection reports have greater impact on restaurant review websites than when they are hidden on the local Council website. In the UK the re-use agenda is now also about improving how public services are delivered and achieving mainstream public policy objectives, in health, education, the environment and so on, as it is supporting the development of the information industries.

10 September 2008
John Sheridan (john.sheridan@opsi.x.gsi.gov.uk), Office of Public Sector Information, UK.

This means government treating the web as:

• a place to provide public services (government to citizen, government to business)

• a platform for data, on which others can build services

At OPSI this approach is called, “Provide and Enable”. It is the core of our own online strategy. Where we provide, we deliver high quality online services direct to the public. Where we enable, others provide services re-using our data. It is thinking that UK Government is adopting more widely and is vitally important for taking forward the re-use agenda forward.

3. IARs as an approach to overcoming barriers to re-use

Government departments and agencies create information assets as part of their public task. These information assets, typically created for internal administrative purposes, have not been developed with onward re-use in mind. Such assets may contain third party intellectual property rights or personal information, for example, which means they cannot easily be released as they stand. The government then needs to make decisions about what information to adapt and make available for re-use. To do this it says to re-users, “tell us what you want, and we’ll tell you if we’ve got it”, to which re-users reply, “tell us what you have, and we’ll tell you what we want”. It was to break this “Catch 22” that Information Asset Registers (IARs) were devised.

An Information Asset Register is a structured metadata catalogue of the key information assets that a public sector organisation holds. In the UK, often it is created alongside a publication scheme (a public sector organisation’s statement about what information it will publish, required under the UK’s Freedom of Information legislation), with the IAR covering primarily unpublished information – but there is no reason for an IAR to be limited to unpublished information alone.

From a re-use perspective the primary purpose of an IAR is for resource discovery – identifying resources that may be available for re-use. They are an important tool for facilitating re-use.

IARs were first introduced in the UK in 1999, to help the information industry and others identify the information assets in government. More recently OPSI re-developed the technology that underpins the current IAR system, with a new platform opening up far more possibilities for the ingest of information into the system through the existing crawler, via a user interface and also by the uploading of spreadsheets or via XML using web services. In addition we are working on support for the OAI-PMH standard.

The quality and scope of IARs has varied across the public sector. The lack of public sector bodies creating and publishing IARs was noted by the Office of Fair Trading as being a potential inhibitor to greater re-use of public sector information, in their Commercial Use of Public Information (CUPI) market study in 2006.

“*6.18 Publication schemes and information asset lists are not sufficient for a potential customer to know all the information held by a PSIH that could be available to businesses. In order for businesses to have this knowledge, the PSIH would also need to produce something like an Information Asset Register (IAR). IARs are part of an initiative across central government which cover virtually all information resources. The initiative is concentrating on listing un-published resources to encourage greater access and re-use. IARs specify rights and restrictions to use of the data. Overall, 81 per cent of respondents to our PSIH survey indicated that they did not have, or had not contributed to an IAR.”*

John Sheridan (john.sheridan@opsi.x.gsi.gov.uk), Office of Public Sector Information, UK.

However it would be foolhardy not to recognise that the emphasis in terms of government action to enable re-use has moved on since the late 1990s. The combination of increase in web publishing and, in the UK, the Freedom of Information regime, has put more and more public sector information in the public domain. The challenges are increasingly around presenting published information in re-usable forms, which is partly a technical challenge in terms of data formats and also finding the right approaches for licensing re-use.

The issue with IARs has been to justify continuing the effort involved in their creation and maintenance. One lesson from the UK is clear: for most officials a more compelling case can be made for IARs in terms of immediate benefit to public sector information holders themselves. An IAR gives an organisation corporate knowledge of what information assets they hold. It is typically of secondary importance to the information holder, that by sharing that information with others, be they other public bodies, private companies or individuals, they can provide a way to let others know what information they may make available for re-use.

Importantly, the UK government's recent Data Handling Review published in June 2008 requires that departments identify their key information assets and information asset owners, as part of a strategic approach to managing information risk. As the issues surrounding data and information management have moved up the agenda in government, so has interest in IARs – not to facilitate re-use, but as a risk management tool.

For those public sector information holders that are part of the Information Fair Trader Scheme (mandatory for Crown bodies that trade in their information), having an asset list is a requirement. In addition OPSI recommends the creation and maintenance of an IAR as being the ideal way for other public sector organisations to fulfil their statutory obligations, under the Re-use of Public Sector Information Regulations to publish "a list of main documents available for re-use".

4. Other approaches to overcoming barriers to re-use

IARs are important to the UK government but are not the only focus for taking the re-use agenda forward. As part of the Power of Information agenda, OPSI has adopted other strategies to help overcome the barriers to re-use, alongside IARs. Where IARs are a supply side solution, our more recent efforts have been focussed on understanding the demand for Public Sector Information. In July 2008 we launched an "Unlocking Service". This provides the government with a way to gather and assess requests for the re-use of public sector information, in a more responsive way. The aim of the service is to give re-users what they want rather than the government tell them what they can have (for example, through an IAR). These issues addressed by the Unlocking Service, are not about access (which are dealt with under access legislation, such as the Freedom of Information Act or Environmental Information Regulations), but all the other pitfalls which can occur when people want to do something with public sector information - copy it, remix it with other data or add value and republish it.

The Unlocking Service works very simply. Someone comes to the service online and describes the public sector information asset they want unlocked for re-use, by posting a request. This describes the problem, what the re-user would like to do with the information and their perfect solution. The request is then posted on the OPSI website. Others can see this request and support it, either by adding a comment or by voting. The idea is the more support a request has, the better the chances of unlocking the information for re-use. OPSI contacts the public sector information holder to see what can be done. If the problem the re-user describes relates to an issue specifically covered by

the Re-use of Public Sector Information Regulations or the Information Fair Trader Scheme, it is treated accordingly - so there is no need to make a separate complaint.

It is a good thing for public sector information holders to have an IAR. It is also good to publish information on the web, not just an IAR entry but the data itself, and better if that information is published in re-usable format and under enabling licensing conditions.

5. IARs and metadata

Information needs to be described in order for it to be useful. That is the role of metadata. However, the nature and type of metadata is changing as the nature of digital content changes. Not all information assets are the same. In fact across the public sector there is enormous variation in types of public sector information holder and in the nature of the information assets they produce.

For example, new APIs and datasets have been released as part of the *Show Us a Better Way* competition, but the approaches for disseminating this data vary considerably, from SOAP and RESTful APIs through to simple spreadsheets. The data may be available, but that does not make it readily accessible on the web, let alone being available using a consistent approach.

In the world of linked data, content becomes more fragmented and more distributed. The distinctions between data and metadata become blurred. Any type of semantic enrichment of data – trying to formally express what data means – involves augmenting information, adding something extra, adding metadata. Such augmentation is not an end in itself, it facilitates re-use of the information.

The view of metadata in the UK is far richer than looking narrowly at bibliographic style descriptions provided by Dublin Core. The descriptions of data will be different, depending on what the information is and how it can be used. In the old world we used to have metadata (e.g. library catalogues or IARs) to aid resource discovery. In the new world, metadata is important for resource discovery, but is capable of delivering much more – for example helping to tailor services around the individual and support personalisation of public services.

Metadata can also enable us to present information in more re-usable forms. An important area is the use of semantic mark-up, embedding metadata at the fragment level, within web pages, to enrich semi-structured and unstructured data. That is why the primary driver of the metadata working group in the UK is to help improve public services – enabled them to be more joined up, not just facilitate better search.

Metadata and metadata standards are key to enabling the large scale re-use of public sector information – but it is not just the metadata of IARs. Yet one of the advantages of an IAR is that it contains consistent metadata for each of the information assets it contains. So why not standardise the metadata elements for an IAR? To have traction, IARs need to work primarily for departments internally requirements, as part of managing risk. Those requirements are likely to differ depending on the nature of the organisation and the information management climate it is operating in.

In the UK, departments are encouraged to publish their IARs on the web. The Office of Public Sector Information offers a web based search facility which allows users to search and access all central government departments individual IARs, through a federated search facility. For such a federated search to work requires some level of standardisation and interoperability between IARs.

There are two approaches to enabling this type of data exchange - the development and adoption of agreed common standards, or, where standardisation is inappropriate, the use of other methods to enable systems to exchange information. This is the realm of semantic interoperability. It is not always necessary to standardise – diversity can be right and good, especially when things are different by nature or requirements divergent.

6. Interoperable IARs made available on the web

It would be possible to devise a standard set of metadata elements for IARs across Europe. However such standardisation approaches need to be approached with great care. Past experience, for example with the application of the UK Government's metadata standard (called eGMS) to websites, teaches us that standardised metadata can lead some public sector organisations to add information simply because it is part of the standard not because they can usefully or meaningfully populate a given element. Such 'box ticking' is wholly counter-productive. In general bad, incorrect or overly generalised metadata is worse then no metadata. This is particularly true of subject classification.

Rather than formally standardise metadata elements in IARs, there should be a common approach for surfacing IARs to the web. What is needed is a pragmatic approach that the majority of public sector organisations can readily use. It is not realistic to expect many public sector information holders to create OAI-PMH compliant metadata repositories. However most can surface their IAR to the web and it is here that a level of interoperability can be achieved. With new Semantic Web standards it is possible to markup textual information inside documents, in effect turning a set of traditional webpages, like an IAR record, into an API. The UK government has been exploring the use of semantic markup inside XHTML documents, using an approach called RDFa, in order to facilitate access, use and re-use of data.

RDFa is a new W3C candidate recommendation which is both easy to understand and fairly easy to implement. Combined with another approach, called GRDDL transformations, it's possible to serve RDFa enabled data in a number of re-usable data formats on the web, such as RDF, XML, JSON and YAML.

The UK government is looking carefully at RDFa and how it might enable re-use, particularly of structured and semi-structured data such as IARs. RDFa is ideal for IARs as existing metadata schemes such as Dublin Core can be easily and explicitly used within it (everyone uses Dublin Core for the "title" element for example). An XHTML template for an IAR record could be produced fairly easily, showing how an IAR record should be marked-up using RDFa.

There are a number of benefits to using RDFa for IARs:

- IARs become more re-usable themselves!
- Existing web based IARs can be 'tweaked' to support RDFa, without creating separate web services so interoperability can be achieved at relatively little cost
- Existing ontology such as Dublin Core can be re-used, but public sector information holders are not entirely constrained. Additional elements can easily be added, by ontology created at European, national or local level.

- Lowly screen-scraping applications benefit from a well-specified and extensible method for embedding semantics, parsers can quickly extract the data whilst more advanced Semantic Web applications benefit from the full power of RDF

- Browser extensions such as Operator and Piggy Bank are able to expose the data to users who are simply browsing the IAR (e.g. dc.coverage can be used to dynamically open a map for the user showing the area specified)

- This approach maintains a close link between semantic information and the provenance of that information

Individual member states, and indeed individual public sector information holders, would be able to use a combination of standard ontologies (such as Dublin Core), and where needed either terms from a common ontology or their own individual ontology.

What is particularly appealing about RDFa from a government perspective is that existing website based IARs can be more or less 'tweaked' to serve RDF data, simply by adding attributes to the mark-up. In principle, it should be as easy to add appropriate attributes to surface the semantics of the data displayed in the browser, so that theses web pages can be parsed and the data simply extracted.

The extent of the tweaking varies depending upon your starting point, but the typical webmaster in a department, agency or local authority knows how to implement a particular mark-up for a type of information. Here then is potentially an easy route for semantically enabling Information Asset Registers and a design pattern for web based, interoperable IARs that others can adopt and use.

7. Conclusions

IARs are an important tool for enabling the re-use of public sector information, but they are not the only tool.

Departments, agencies and other public sector organisations are more likely to invest in creating and maintaining an IAR as part of their information risk management then they are to facilitate re-use. That means that IARs need to fulfil an internal business needs first and foremost. Consequently, standardising metadata elements for IARs from the re-use perspective alone could be counter-productive, as IARs need to meet internal requirements.

From a re-use perspective the emphasis needs to be, not on the elements that make up an IAR, but on ensuring IARs are published to the web in an interoperable way.

One pragmatic approach for achieving some degree of interoperability for IARs would be to use RDFa, so that the data contained in the web pages that show IAR records can be readily identified and extracted, in effect turning IAR web pages into an IAR API. Alongside this there needs to be some mechanism for discovering IAR records. The sitemaps protocol and ATOM feeds provide two good candidate technologies for doing this in a web centric way.

Whilst sceptical about the merits of fully standardising the metadata elements for IARs across Europe, the UK Government is keen to contribute to work to harmonise interoperable IARs, made available on the web, in a way that maximises their potential for re-use, to aid resource discovery and to support re-users.

John Sheridan (john.sheridan@opsi.x.gsi.gov.uk), Office of Public Sector Information, UK.