LINKED DATA REGISTRY: A NEW APPROACH TO TECHNICAL REGISTRIES

Maïté Braud, Pauline Sinclair, and Robert Sharpe
Preservica Ltd, 26, The Quadrant, Abingdon Science Park, Abingdon OX14 3YS, UK

Abstract

Technical Registries are used in digital preservation to enable organizations to maintain definitions of the formats, format properties, software, migration pathways etc. needed to preserve content over the long term. There have been a number of initiatives to produce technical registries leading to the development of, for example, PRONOM, UDFR and the Planets Core Registry.

However, these have all been subject to some criticisms. One problem is that either the information model is fixed and difficult to evolve or flexible but hard for users to understand. However, the main problem is the governance of the information in the registry. This has often been restricted to the host organization, which may have limitations on the investment they can make. This restriction has meant that, whilst other organizations have, perhaps, been free to use the registry they have been unable to add to or edit the information within it. The hosts of the registries have generally been receptive to requests for additions and change but this has still led to issues with timing or when different organizations cannot agree (or just utilise or interpret things in different ways).

In this paper we describe a new approach, which has used linked data technology to create the Linked Data Registry (LDR). This approach means it is simple to extend the data model and to link to other sources that provide a more rounded description of an entity. In addition, every effort has been made to ensure there is a simple user interface so that users can easily find and understand the information contained in the registry.

This paper describes what is believed to be the first linked data technical registry that can be deployed widely. The key element of the new approach is the distributed maintenance model which is designed to resolve the governance problem. Any organization hosting an LDR instance is free to add and edit content and to extend the model. If an instance of LDR is exposed on the internet, then any other organization is free to retrieve this additional information and hold it in its own LDR instance, alongside locally maintained information and information retrieved from other sources. This means a peer-to-peer network is established where each registry instance in the network chooses which other registry instances to trust and thereby from whom to receive which content. This gives control to each individual organization, since they are not dependent on anyone else but can choose to take different content from appropriate authoritative sources. At the same time it allows collaboration to reduce the administrative burden associated with the maintenance of all of the information.

Introduction

Role of Technical Registries

One of the key threats to the preservation of digital material is that “Users may be unable to understand or use the data, e.g., the semantics, format, processes or algorithms involved” (Kuipers, 2009).

This issue is addressed in the OAIS model through the development of Representation Information networks (CCSDS, 2012). Some of this might be specific to a given Information Object (e.g., data from a one-off experiment might need to record information related to the instrument calibration and quality control that took place) or it might apply very commonly (e.g., the need to understand the specification of PDF/A). This means that Representation Information networks will consist of some information maintained locally (to hold information specific to the Information Objects held in that repository) and some information that is probably best maintained remotely from the repository (or at least it can be done more efficiently, e.g., not every organization using PDF/A needs to be an expert in the details of its specification).

The need for Representation Information networks is well established in data-holding institutions. This is because, for example, data gathering often utilises new combinations of techniques, methods and algorithms and thus, in order to be able to understand the results, a repository needs to be able to reference information related to these and yet does not necessarily want to repeat this information with every data set.

In memory institutions traditionally the problem has been handled in different ways using different terminology but conceptually it is the same approach. For example, usually such institutions create a catalogue entry to describe (at least at a high level) each record it holds. This catalogue entry, as well as describing information specifically about the record, may reference other information (e.g., a description of the collection to which the record belongs, or links to other controlled sources such as organizations, people or events related to the record). These controlled sources are then described in turn (externally to the individual catalogue entry) and may, themselves, reference another external source. This creates a network of information that helps a user to understand the semantics of a record.

For example, imagine a genealogist looking at the history of an ancestor. From the records of a national archive, they might be able to find out that their ancestor was in the army and served in a given regiment between two dates. The national archive might maintain a separate list of information about every regiment in the national army but it might not contain detailed information about each regiment, such as where that regiment was posted on a given date. However, this information might be available from a regimental museum. Hence, a given user (with sufficient knowledge and skill) can find out where their ancestor was posted on a given date through the use of a network of representation information that will involve information held with the record, information explicitly linked to the record and information implicitly linked to the record.

For memory institutions, this sort of network applies to paper records as well as digital records and they have been in existence for some time. The advent of digital technology has made catalogues of information easier to maintain, more accessible, easier to search and easier to link to each other but the fundamental information storage and retrieval process has not changed. However, the advent of digital information has led to new problems such as the ability to continue to interpret, for example, a file of a specific format that constitutes all or part of the original record.

To solve this problem various attempts have been made to add such information to the existing, relevant representation information networks. This has included the development of 'Technical Registries' which are designed to be repositories of key facts about things that are important to the environment needed to interpret digital records and/or the environment needed to preserve such records.

There have been a number of high profile attempts to create such a registry including PRONOM (http://www.nationalarchives.gov.uk/PRONOM/Default.aspx), UDFR (http://www.udfr.org) and the Planets Core Registry (http://www.openplanetsfoundation.org/planets-core-registry). These registries have provided significant advantages and at least some of them are in regular use. PRONOM, for example, is used as the basis for the format signatures that underpin the widely-used file format identification tool, DROID (http://digital-preservation.github.io/droid), while the Planets Core Registry has been used as the basis for automated characterisation and migration decisions within Preservica's (part of the Tessella group) digital preservation systems: Preservica EE (formerly known as SDB) and Preservica CE (http://preservica.com).

Other initiatives such as the "Solve the File Format Problem" (http://fileformats.archiveteam.org/) or the Community Owned digital Preservation Tool Registry (COPTR) (http://coptr.digipres.org/) have already demonstrated the benefit of using crowd sourcing to collate information relevant to the Digital Preservation community but these repositories do not offer machine-to-machine interfaces and thus are aimed mainly at researchers or manual curation.

Limitations of current registries

However, all of these registry initiatives have also been subject to two main criticisms.

The first is that the set of entities modelled, the properties held about such entities and their relationship to other entities has been hard to expand and/or hard to interact with. Either of these issues makes it hard to integrate this information as part of a representation information network. For example, it would be desirable to be able to link a locally-held record about a format to, say, its formal specification. In some existing registries this could be done by, say, uploading a copy to the Technical Registry but then this would not be updated if some error was found in the specification and updated on, say, the official website.

There have been two contrasting approaches to this issue of expandability and usability. The first has been to use a fixed-schema database with a user interface intricately linked to that schema. This approach (used in PRONOM and the Planets Core Registry) makes the system easy to use but hard to expand. The alternative approach (used in UDFR) has been to use a linked data approach which is easier to expand. However, linked data is a technology designed for computer-to-computer interactions, meaning that it can be hard for non-technical users to interact with the information. UDFR has made some effort to create a user interface to help with this but arguably it is harder to use the software to find information than, for example, in the fixed-schema, harder-to-expand PRONOM system.

The issue has already been raised in previous papers, and initiatives such as the P2-Registry (Tarrant, 2011) recognised and proved the benefit of the Linked Data approach while highlighting that exposing SPARL query interfaces directly to end users might be too complex for a lot of people to use.

The second issue is one of governance of the information. Since these registries have been used by organizations other than their hosts, there have been issues about what to do when information is incomplete, in error or possibly subject to just being an opinion. For example, some organizations have wanted to extend the range of formats that is covered by PRONOM. The UK National Archives (the hosts of PRONOM) have been as proactive as possible at supporting such requests but the need for them to go through appropriate checks and their limited resources means that it can take some time before a request leads to a registry update. In addition, there have also been cases where there have been disagreements within the community about format definitions, and cases where an information update has changed existing behaviour causing systems that relied on the previous behaviour to stop working as expected.

New Approach

This paper describes a new type of Technical Registry designed to solve these problems: the Linked Data Registry (LDR). Like UDFR it uses linked data technology (http://linkeddata.org/), which allows flexible linking of resources to other resources thereby offering a solution to the expandability part of the first issue.

In addition the registry aims to be as easy to search, and to view and edit entities as a fixed-schema system. This means it also offers a solution to the usability part of the first issue. Searches of linked data systems use a search language called SPARQL that is conceptually similar to the structured query language (SQL) used by more traditional relational databases. In many linked data systems a SPARQL end point is considered sufficient to allow for searching, viewing and editing of content. However, the users of a Registry should not be assumed to be sufficiently technically savvy to write queries using SPARQL or to be able to interpret the raw results, any more than users of a traditional relational database would be expected to write SQL statements or interpret the raw results this would produce. Creating a method of allowing searching, viewing and editing of linked-data information in a manner that is natural to non-technical users is a non-trivial issue that has been the subject of considerable research effort (Davies, 2010). In this paper we describe how we have attempted to solve this problem. It is inevitably a design compromise but one that we believe is optimized to balance expandability and ease of use.

Crucially, LDR also addresses the issue of governance. It allows a network of registries to be created that can be replicated peer-to-peer, thereby removing the need for any organization to be dependent on any other for the maintenance of information, unless it chooses to be so.

Linked Data

Linked data is becoming a more commonly used technology but some readers may be unfamiliar with it or unclear what terminologies such as resource, subject, predicate and object mean. This section provides a very brief introduction, which should be sufficient to understand the rest of this paper.

A resource is the linked data term for an entity; examples include file format, software and migration pathway. A resource needs to be identified uniquely by a URI (Uniform Resource Identifier).

A resource is described by a set of statements (expressed as subject - predicate - object). Statements can be either simple or complex:

- A simple statement is a statement where the object is of a simple type: e.g., a String or an Integer, but crucially *not* another resource.
- A complex statement is a statement where the object is another resource.

For example:

- “Resource A” “has MIME type” “image/jpeg”
- “Resource A” “has PUID” “fmt/44”
- “Resource A” “has extension” “JPEG”
- “Resource A” “has extension” “JPG”
- “Resource A” “has version” “1.02”

are all simple statements in the form subject - predicate - object that describe and identify resource A (aka JPEG file format v1.02).

Resource A “has internal signature” Resource B (where resource A is a file format and resource B is a DROID internal signature) is an example of a complex statement. In this case the DROID internal signature object will itself be an agglomeration of statements that define and describe it.

Information Modelled

In this first version of LDR the information modelled needed to be sufficient to allow efficient (and automated) preservation-related activities to take place. However, after meeting this sufficiency criterion, the data model has been minimised deliberately.

This was done partly to keep the problem tractable but also partly based on the experience of developing the Planets Core Registry. In that project we found that there was a wish to expand the data model to include every attribute that might possibly be needed in the future. This was understandable since the technology used (a relational database with a fixed graphical user interface) meant that it was hard to expand the system after it was initially completed. However, this meant in practice that large tracts of the data model were left unpopulated. Perhaps worse was that it was not clear if the lack of information meant that the data model was not useful, the information was not valuable enough to be collected, the information was too hard to collect, or maybe it had not been collected yet.

Hence, in this study, it was decided to use a technology that was much easier to expand (linked data) and to start out by only modelling the information that was known to be of interest (essentially the entities that were populated in the Planets Core Registry).

These entities could be split into two classes: factual information (information that could reasonably be expected to be held in common by lots of agencies without controversy) and policy information (information about what to do when, that might be relevant to only one repository). In LDR these two classes of information are held separately, but still linked. It should be emphasized that this is not a hard and fast distinction: just a pragmatic one. Hence, it is possible for organizations to disagree about information (such as the exact definition of a format) while it is also possible for organizations to share policies. The use of a peer-to-peer network (see the "Replication" section) allows both of these cases to be covered.

Factual Information

The Linked Data Registry models a number of key factual entities aggregated into five groups:

- File formats (with associated DROID internal signature and byte sequences)
- Software
- Related software tools (including the tool's purpose and parameters)
- Migration pathways
- Properties and property groups

The decision to create these five groups of entities was based on how these entities are used by users. For example, a user would naturally view, create or edit information about a format and then expect to add or create an internal signature for that format. Linked data concepts mean that this relationship could be considered the other way around (i.e. internal signatures are associated with formats) especially given that a single internal signature is often associated with multiple formats. However, humans tend to look up the signatures associated with formats more often than the other way round and would tend to add new signatures based off information derived from a format's specification.

This aggregation is important for the user interface needed to interact with the system (see the "Search, View, and Edit" section below). It is less important from a technical perspective which can safely consider the resources to be linked to each other from any perspective. The impact of this aggregation on the expandability of the model is discussed in the "Expansion" section below.

Each of these five groups of entities is discussed in turn.

Format Information

This entity group models file formats, including internal signatures and the byte sequences of internal signatures. It is based on the model established by the UK National Archives as part of their Linked Data PRONOM research project

(http://labs.nationalarchives.gov.uk/wordpress/index.php/2011/01/linked-data-and-pronom).

Attribute	**Repeatable?**	**Link to other Resource**
Identifier	N	N/A
Name	N	N/A
Version	N	N/A
Description	N	N/A
Release Date	N	N/A
Withdrawn Date	N	N/A
Internet Media Type	Y	N/A
File Extension	Y	N/A
Has Internal Signature	Y	Internal Signature
Is Rendered By	Y	Software
Is Created By	Y	Software
Is Validated By	Y	Software Tool
Has Properties Extracted By	Y	Software Tool
Has Embedded Objects Extracted By	Y	Software Tool

Attribute	Repeatable?	Link to other Resource
Has Property	Y	Property
Belongs To Format Group	Y	Format Group
Has Priority Over	Y	File Format
Has Lower Priority Than	Y	File Format
Is Previous Version Of	Y	File Format
Is Subsequent Version Of	Y	File Format

Table 1: File Format Attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Description	N	N/A
Has Byte Sequence	Y	Byte Sequence

Table 2: Internal Signature Attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Position	N	N/A
Sequence	N	N/A
Byte Order	N	N/A
Offset	N	N/A
Max Offset	N	N/A

Table 3: Byte Sequence Attributes

Software Information

This entity group simply models the existence of a software package.

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Version	N	N/A
Description	N	N/A
Release Date	N	N/A
Withdrawn Date	N	N/A
Vendor	N	N/A
Licence	N	N/A
Web site	N	N/A

Table 4: Software Attributes

Tool Information

This entity group models the use of a piece of software as a tool for characterisation, migration, or some other purpose. It allows modules of software packages to be specified and classified.

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Implementation Details	N	N/A
Has Purpose	Y	Tool Purpose
Has Tool Parameter	Y	Tool Parameter
Belongs To Software	N	Software Tool
Can Extract Property	Y	Property

Table 5: Tool Attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Applies To File Format	N	File Format
Has Priority	N	N/A

Table 6: Tool Purpose Attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Value	N	N/A

Table 7: Tool Parameter Attributes

Migration Pathway Information

This entity group models a migration pathway and its roles (e.g. presentation, preservation) and uses.

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Has Source Format	N	File Format
Has Target Format	N	File Format
Uses Tool	N	Tool
Has Target Format Group	N	Format Group
Has Validation	Y	Migration Pathway Validation
Has Role	Y	Migration Pathway Role

Table 8: Migration Pathway Attributes

Property Group Information

This entity group models a ‘Property Group’, which is a type of information object (e.g., document, video, website, etc.), the properties that might be expected to be measured for each such property group, and the groups of formats in which this might be manifested (called ‘Format Groups’). For example, a property group called ‘Image’ might have a series of properties (e.g., height, width, colour space, etc.) and be manifested in a whole series of ways (e.g., as a part of the TIFF format group, as a part of the JPEG format group etc.).

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Has Property	Y	Property

Table 9: Property Group Attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A

Table 10: Property Attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Belongs to Property Group	N	Property Group

Table 11: Format Group Attributes

Policy Information

LDR can also model policy information. In this first version this is restricted to two simple policies.

Tool Priority

This can be used when multiple tools are present in the Registry to carry out a task (e.g. format validation) to determine which should be used in preference to the other(s). Tool Priority is described in Tool Purpose (see Table 6) but is part of the policy section of data.

Migration Pathway Validation

This can be used to determine which properties should be measured before and after migration and compared in order to check that significant properties have been maintained acceptably. It allows a tolerance to be set for cases where the value of the significant property is allowed to change (within limits) during migration.

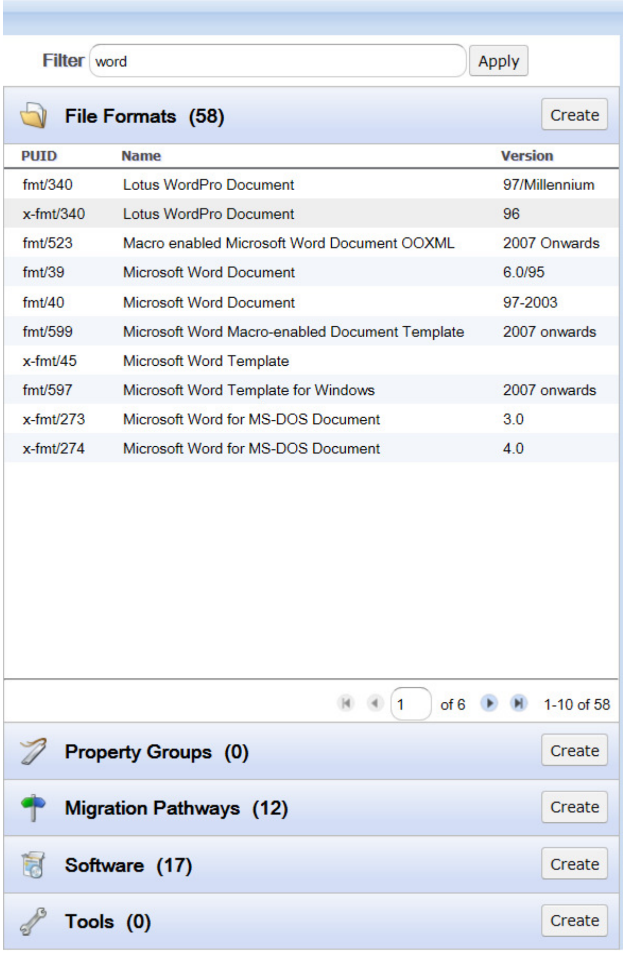
Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Source Property	N	Property
Target Property	N	Property
Tolerance	N	N/A

Table 12: Migration Pathway Validation Attributes

Using the Registry

Search, View, and Edit

As described above, one of the key features of LDR is that the registry has an easy-to-use user interface. This allows users to search for and view information about each currently-supported entity. Also, users with the appropriate authority can use this interface to edit information about an entity and/or add a new entity. Very importantly there is no need to understand linked data concepts or how the information is organised and stored in order to use this user interface.



Filter word Apply

File Formats (58) Create

PUID	Name	Version
fmt/340	Lotus WordPro Document	97/Millennium
x-fmt/340	Lotus WordPro Document	96
fmt/523	Macro enabled Microsoft Word Document OOXML	2007 Onwards
fmt/39	Microsoft Word Document	6.0/95
fmt/40	Microsoft Word Document	97-2003
fmt/599	Microsoft Word Macro-enabled Document Template	2007 onwards
x-fmt/45	Microsoft Word Template	
fmt/597	Microsoft Word Template for Windows	2007 onwards
x-fmt/273	Microsoft Word for MS-DOS Document	3.0
x-fmt/274	Microsoft Word for MS-DOS Document	4.0

1 of 6 1-10 of 58

Property Groups (0) Create

Migration Pathways (12) Create

Software (17) Create

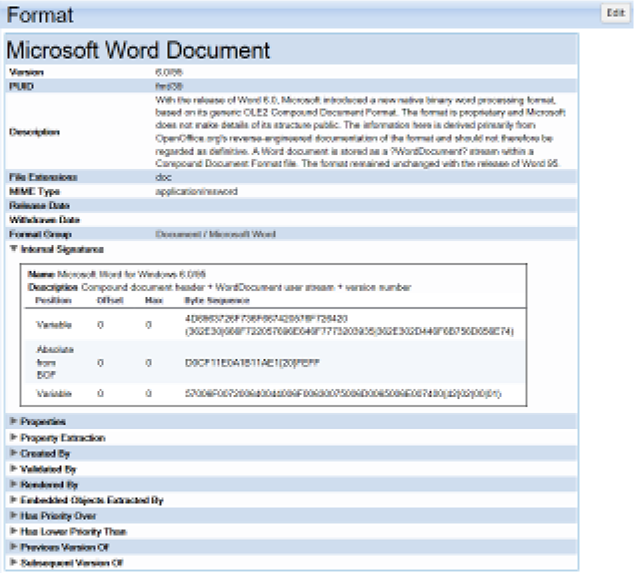
Tools (0) Create

Figure 1: Simple-to-use search in the registry

This usability is achieved by using a single user interface form for each of the five aggregations of factual information described above (format information, software information, tool information, migration pathway information and property group information). For ease of use, the policy information is superimposed on these forms (so tool priority is displayed with the software tool entries and migration pathway priorities with the migration pathways).

Rather than provide a complicated search interface, LDR allows users to filter the lists of entities in each of the five aggregations. There is a single filter box (see Figure 1) that filters the entity lists as each letter is typed. This makes it easy for users without training to find the information that they wish to see.

Once the user has located the information they are looking for in the relevant category, simply clicking on the item will display the information available to them. Initially the key information (e.g., name, version, identifier etc.) is shown. More detailed information (such as the internal signatures of a format, including the list of byte sequences of each such internal signature) can be displayed as desired (see Figure 2).



Microsoft Word Document

Version	6.0/95
PUID	fmt/39
Description	With the release of Word 6.0, Microsoft introduced a new native binary word processing format, based on its generic OLE2 Compound Document Format. The format is proprietary and Microsoft does not make details of its structure public. The information here is derived primarily from OpenOffice.org's reverse-engineered documentation of the format and should not therefore be regarded as definitive. A Word document is stored as a ?WordDocument? stream within a Compound Document Format file. The format remained unchanged with the release of Word 95.
File Extensions	doc
MIME Type	application/msword
Release Date	
Withdrawn Date	
Format Group	Document / Microsoft Word

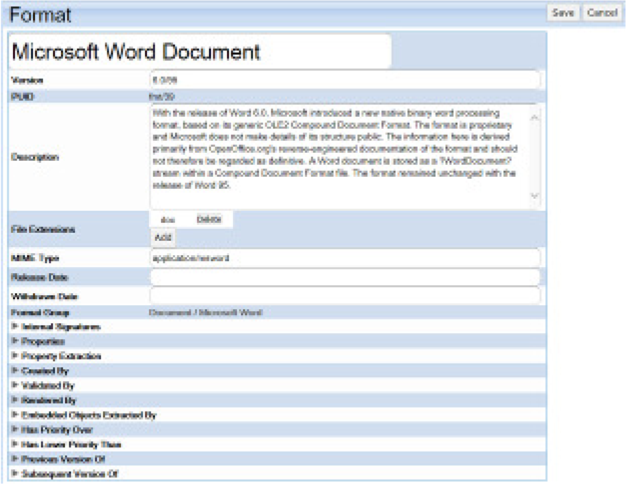
▼ Internal Signatures

Name Microsoft Word for Windows 6.0/95
Description Compound document header + WordDocument user stream + version number

Position	Offset	Max	Byte Sequence
Variable	0	0	4D6963726F736F667420576F726420 (302E30|068F72205769 6E646F7773203935|302E302D446F6B756D656E74)
Absolute from BOF	0	0	D0CF11E0A1B11AE1{20}FEFF
Variable	0	0	57006F007200640044006F00630075006D0065006E007400{2}{02}{00}{01}

Figure 2: Easily-understandable format information

If a user has sufficient authority to be allowed to edit information, then they can access an editable version of the user interface. This allows text to be edited, new items to be created etc. If as part of editing, a link to another resource needs to be created, then the user can choose to link to an existing resource and/or add a new resource as appropriate.



Microsoft Word Document

Version	6.0/95
PUID	fmt/39
Description	With the release of Word 6.0, Microsoft introduced a new native binary word processing format, based on its generic OLE2 Compound Document Format. The format is proprietary and Microsoft does not make details of its structure public. The information here is derived primarily from OpenOffice.org's reverse-engineered documentation of the format and should not therefore be regarded as definitive. A Word document is stored as a ?WordDocument? stream within a Compound Document Format file. The format remained unchanged with the release of Word 95.
File Extensions	doc Delete Add
MIME Type	application/msword
Release Date	
Withdrawn Date	
Format Group	Document / Microsoft Word

Figure 3: Editing format information

Each entity created by an organization will have a globally unique resource identifier (of the form: http://Creating_Organisaiton_Name/Entity_Type/Locally_Unique_Identifier).

Audit Trail

A record of every change to every resource (including its initial creation) is maintained in an audit trail. This is sufficient to allow changes to be reversed.

However, the most important aspect of the audit trail is to be able to determine which entities have been added or edited since a certain point in time. This allows different entities in the network of registries to be replicated, knowing what has changed since the last such replication. The replication process is discussed in more detail in the section below on replication.

Automation

LDR can also support key digital preservation automation features.

The first of these is the creation (and export) of a DROID signature file. This is important since it allows any organization not only to add its own formats, and their signatures, but also to be able to use DROID to identify them, even if the UK National Archives (who control the globally-available DROID signature file) choose not to add them to their registry.

In addition, LDR also comes with the machine-to-machine interfaces needed to allow a digital preservation system to query it automatically and thereby drive decisions relating to characterisation, preservation planning, and preservation actions such as migration. The adequacy of this interface has been demonstrated by using it to automate preservation-related activities within Preservica's (part of the Tessella group) digital preservation systems, Preservica EE (formerly known as SDB) and Preservica CE (http://preservica.com). This demonstrates that it is an adequate replacement for the less flexible, existing Registry previously used for this purpose (the Planets Core Registry).

Expansion

The initial set of modelled entities included in LDR has been limited deliberately to those commonly used in digital preservation systems. Some of the existing registries support a wider data model but, as discussed above, these entities have not been heavily populated with data (if at all).

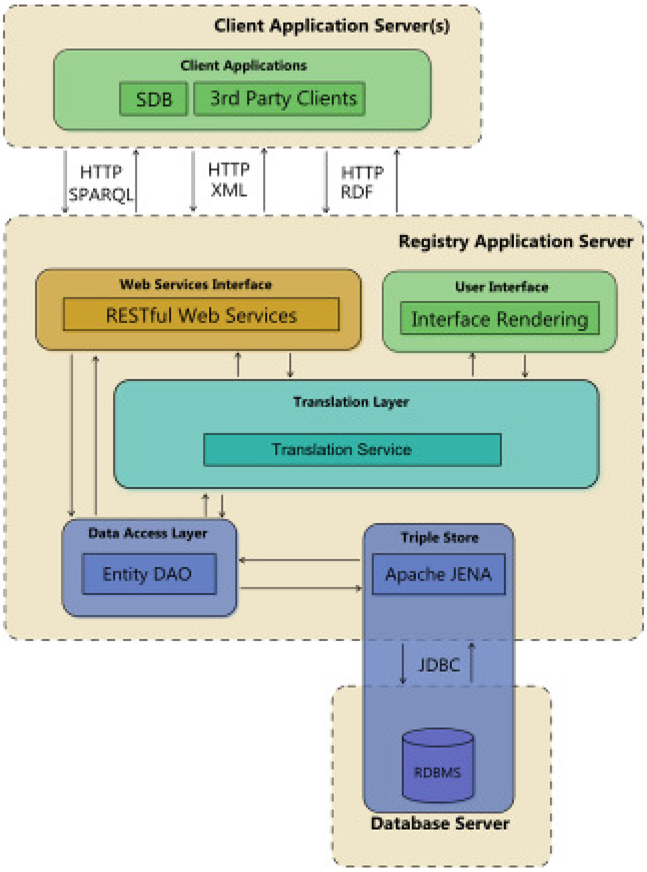


Figure 4: LDR architecture

Since the new registry utilises linked data technologies, it is easy to add resources to LDR and/or link to an external source to expand this model, if necessary. This expansion could be an additional property of an existing entity or it could be the addition of a completely new type of complex entity.

When the data model is expanded the user interface can also be expanded but, since the user interface is not created dynamically from the data model, this will take more effort. It would be possible to design a generic user interface but this would not meet one of the aims of this system: to ensure that users can easily see information and, where appropriate, add new information and edit existing information in ways that can be readily understood. We felt that a generic interface would be a big barrier to this. Hence, this is a design compromise.

LDR has been designed to offer a number of options for dealing with expansion because of this need to make a design compromise (see Figure 4). At the core of the system is a triple store with an exposed SPARQL interface. To offer a more advanced interface to client applications, there is a translation layer that combines multiple triples into more convenient to use data objects that can be accessed by such clients as either XML or RDF aggregations. The Registry user interface itself consumes these aggregations and displays the information.

Hence, the options for adding new entities are (in increasing degree of effort):

- The simplest option is just to add entities to the triple store. These will be available for access by client applications via SPARQL queries and RDF.
- The next option, in addition to adding the entities to the triple store, is to enable the translation service so that the aggregations in XML can be created and validated against their XML schema. These will be available for access by client applications via a RESTful web service interface.
- The most complete option is to update the user interface as well, so that the additional information is displayed here. This could be adding additional aggregations or adding to the existing ones.

In the first version of LDR all entities in the triple store are aggregated in the translation layer and most are displayed in the user interface. It is possible that future versions will be expanded without doing this (i.e. the user interface might best be seen as a filtered view of the total information held in the triple store). It is certainly important that expansion is not prevented by the need to expand all the architectural layers.

This is an interesting design compromise that only time will tell if it has been optimised appropriately.

Replication

Network of Nodes

LDR is designed to be used as a network of registry instances, or nodes, with each node in the network being able to control its own factual information. Clearly maintaining all this information is a potentially large burden. Hence, each node can choose which node(s) to extract content (or a subset of content) from. The audit trail allows the set of potential updates since the last such extraction from a target node to be identified easily.

This means that every node can independently choose who to trust about what (and what information it wants to take on the responsibility for maintaining itself). It also means that different nodes can choose to maintain (and publish) different subsets of the total information space. These subsets can overlap with other nodes since it is up to each other node in the network to choose which other nodes to trust for which content. It does not necessarily matter if different nodes in the network hold different information about nominally the same entity, provided that the information used is appropriate to that community.

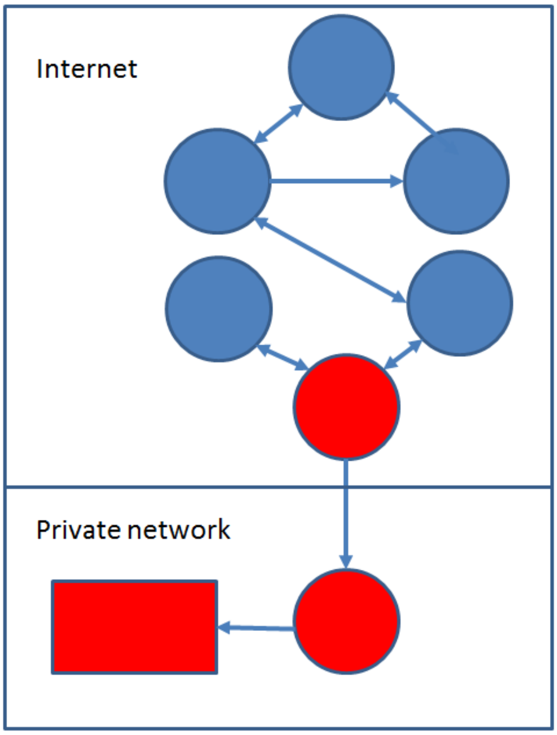


Figure 5: Possible network of nodes. Each Registry node is a circle with a rectangle representing a repository. A single organisation controls the elements in red, while blue entities are from different organisations.

Hence, LDR uses a peer-to-peer replication model. The advantage of this over alternative network configurations (such as ball-and-spoke, where one central node controls the content) is that it removes the need for centralised governance. Each node can control its own information and, if it chooses to, update that information immediately. At the same time the ability to extract content from other nodes means that the burden of maintaining information can be shared.

Figure 5 shows how this network could be used. In the top part of the diagram are a series of Registry nodes (each represented as a circle) in the internet which have chosen to trust all or part of the information maintained in other nodes. One organization (shown in red) is a part of this network but operates its production repository (the rectangle) inside a private network protected by a firewall. A separate (private) Registry instance serves the repository and is updated only from that organization's public Registry instance in a controlled manner. To enable this scenario, LDR supports a data dump to enable replication without the need for a network link between nodes.

Shared Instances

It is also possible for multiple organizations to share a registry instance. This allows for instance-level factual information, which would normally be controlled by a host organization (through a combination of local maintenance and choosing which other instances to trust). However, each organization using the instance could set their own independent policy information whilst sharing factual information.

Future Work

LDR is being rolled out first to Preservica's customer base but then will be offered more widely. If there is sufficient interest a community version could be created.

It does bring a number of interesting challenges. It removes the need for central governance but this does not mean that there should not be guidelines for updating and adding new entities. There are likely to remain islands of excellence on which lots of other organizations will choose to depend (e.g., organizations might rely on the UK National Archives for information on standard formats, as many do already via PRONOM; customers of commercial repository supplies might rely on the provider of this software for much of the information about the available tools and migration pathways used in their software, etc.). It will be interesting to see who organizations choose to trust for which subsets of information and on what basis. It will also be interesting to see how organizations choose to take on the burden of the maintenance of some subsets of the necessary information themselves.

To avoid editing conflicts and the creation of competing linked-data resources, Preservica has proposed a protocol to define who should update what and what the procedure is should another organization want to make an update. We hope that this protocol will be adopted by the initial users of LDR (our customer base). Such a protocol is needed in order to build trust in the network, so that it becomes a network in truth (which collaborates on building and maintaining technical Representation Information), rather than a set of islands of local information (with each organisation having to invest lots of effort to maintain all of its own data).

In practice the protocol is a set of 'etiquette rules' that organisations agree to abide by, so as to bring order to the process of creating new linked-data resources and editing existing ones. Organisations can appoint themselves as 'experts' for a particular type of resource (their area of interest). If an organisation wishes to create or update a resource in another organisation's area of interest, it is expected to give the 'expert' organisation the opportunity to make the change itself, with a deadline for a response. Only if the 'expert' organisation is unable to help within the deadline can the initial organisation make the change; the 'expert' organisation can choose to endorse the change retrospectively. Where areas of interest overlap, the organisations are expected to agree between themselves as to how they will manage potential conflicts. We hope that these rules will encourage LDR's users to maintain the information in it proactively, and make it easy for them to do so.

In addition, it will be interesting to see how the data model is expanded over time. We would anticipate an increase in the use of links to expand the model by linking to existing, external linked-data models as opposed to adding complex new entities to the system.

Conclusions

Technical Registries (used to help with the preservation of digital documents, images and related content) are part of a continuum of representation information networks that include other forms of digital content and non-digital content. Some parts of this network have existed for centuries whilst others (including those covered by technical registries) are new and currently incomplete. The key lessons of existing technical registries are that:

- They must be expandable and must be able to be linked to other parts of this network.
- They must be easy to use without detailed technical knowledge.
- There must be local control of governance.

This paper describes what is believed to be the first linked-data technical registry that can be deployed widely, thereby allowing the creation of a network of information maintained by a diverse and (loosely) collaborating community.

This registry has balanced the need to expand the data model with the need to make the entities in that data model easily findable, viewable and editable by non-technical users.

It establishes a replication and governance model for this network based on a peer-to-peer approach. This allows each organization to choose who to trust and which information to maintain itself. Time will tell how this new ability is utilised.

Acknowledgments

The research leading to these results was performed as part of the ENSURE project and has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270000.

Bibliography

Kuipers, Tom, and van der Hoeven, Jeffrey (2009). “Insight into Digital Preservation of Research Output in Europe” <http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf> [Checked 28/08/2014].

The Consultative Committee for Space Data Systems (2012). “Reference Model for an Open Archival Information System (OAIS)”. CCSDS 650.0-M-2. Magenta Book. <http://public.ccsds.org/publications/archive/650x0m2.pdf> [Checked 28/08/2014].

TARRANT, David, HITCHCOCK, Steve, and CARR, Les (2011). “Where the Semantic Web and Web 2.0 Meet Format Risk Management: P2 Registry”. The International Journal of Digital Curation, Vol. 6, No. 1, p. 165-182. DOI: 10.2218/ijdc.v6i1.180, online: http://www.ijdc.net/index.php/ijdc/article/viewFile/171/239

DAVIES, Stephen, HATFIELD, Jesse, DONAHER, Chri, and ZEITZ, Jessica (2010). “User Interface Design Considerations for Linked Data Authoring Environments”. In: Proceedings of the Linked Data on the Web Workshop (LDOW2010), Raleigh, North Carolina, USA, April 27, 2010, CEUR Workshop Proceedings, ISSN 1613-0073, online: http://ceur-ws.org/Vol-628/ldow2010_paper17.pdf.