# Review of National Libraries Global

*Bill Oldroyd, Technical Consultant to The European Library*

## 1.    Introduction

This is a review of the prototype system developed by the National Library of New Zealand to support the "National Libraries Global" service.

## 2.1    Technical Review - Infrastructure Components

The NLG infrastructure broadly comprises 4 components:

- A harvester

- A search engine

- A mechanism for including thumbnails

- A web-server

This infrastructure is mainly based on a standard architecture which uses components that match the choice made by many other systems with similar objectives. For example, they form a very good match with services such as The European Library (TEL) and Europeana.

| Component | TEL | Europeana | NLG |
|---|---|---|---|
| Search engine | Lucene | Lucene | Lucene |
| Search application layer | SOLR | SOLR | SOLR |
| Web server | Apache / Tomcat | Apache / Tomcat | Apache / Tomcat |
| Database | Postgres (migrating to Lucene) | | MySQL |
| Programming language | Java | Java | Java |

However, all 3 services use a bespoke component for the harvesting of records for input to the database, and to some extent NLG has different and more comprehensive functionality for the harvesting process. (The NLG harvester harvests directly from web pages as well as using the standard harvesting protocols.)

For TEL, the TELplus project included the development of an open source package called REPOX which, whilst only harvesting using OAI PMH and Z39.50, sets out to automate the regular harvesting of metadata records and the associated metadata conversion. REPOX also allows the aggregation of the harvested records, so that it may function as a data provider to a more centralised system. For example, TEL uses REPOX to aggregate records from libraries and then supply them to Europeana as a data provider. Automating the harvesting processes is a key function of REPOX, as the experience of TEL in harvesting a large number of collections manually shows it to be a time-consuming and tedious manual activity and to some extent prone to mistakes.

Author: Bill Oldroyd
20th May 2010

Harvesting of metadata based on standard metadata formats and protocols does not normally allow the harvesting of the associated image thumbnails (or other summary objects). An additional function is required to harvest the associated thumbnails or other digital objects. The three services adopt different approaches :

- TEL draws thumbnail images from the the metadata providers' services,

- Europeana harvests the thumbnails and operates a dedicated thumbnail server,

- NLG redirects http to the host service using a proxy.

The approach taken by Europeana gives the greatest guarantee that service requirements can be met. In the other cases, when thumbnails are fetched from other services there is a chance those services may not be accessible for some reason or there may a change in the network address used to access the thumbnails. (The problems are illustrated by an error in access to the digital objects from the Singapore National Library, where the links to the objects are clearly not the ones described by the metadata.) The thumbnails may fail to be displayed.

On the basis that NLG utilises Open Source Software, as opposed to commercial package software, the architecture can be regarded as state of the art. And, although it is open source software, the functionality used by NLG is equal to commercial software. The various software components are supported by dynamic open source communities actively developing and supporting the software more or less in line with the requirements of these services.

The risks that the software will not be maintained in the long term or that the functionality will not be developed further are small, at least for the foreseeable future.

## 2.2    Technical Review - Scalability

On the basis that NLG will load metadata records for individual objects, NLG has the potential to require a very large database. TEL has the long-term target of indexing 150 million records and Europeana at least 40 million records, but neither has reached this target yet. TEL currently has more than 20 million records (this includes bibliographic records as well as those for digital objects) and Europeana more than 7 million records.

NLG also has the potential for a very large number of concurrent users, especially in conjunction with specific events such as project launch, system upgrades or promotions. It is extremely difficult to estimate the number of users likely to use NLG.

However, it is possible to consider the capability of the technical solution to see if it has the possibility of scaling to a very large load.

Scalability is affected by factors such as the limitations of software, the required level of processing, data storage performance and associated manual processes supporting the service. For example, it includes the following :

- Overall capability of the search engine

- Data load

- Harvesting workload

2

The search engine, Lucene, provides implementation strategies (index sharing and replication) which will allow the service to scale across processors, memory and storage if these have limitations caused by either the amount of data or concurrent user loads. Given that the NLG data volumes will build up over time, for several years the service is likely to remain within the proven capabilities of the search engine to handle large volumes of data and users. There will be the additional costs of providing the extra hardware and support to meet growing demand.

However, the scalability of the service also depends on the harvesting process and associated indexing which will be impacted by the number of separate collections to be harvested and the frequency of updates to the index. The experience of TEL, which currently includes over 400 collections, indicates the importance of an efficient and automatic approach to harvesting. Otherwise this can become a significant manual activity, especially if this involves dealing with the quality of metadata supplied by partners. Europeana experience suggests the importance of rejecting poor quality metadata.

Processing capacity may also be unable to support the large volumes of use caused by well publicised and high profile launches. Unless a loss of service is acceptable, a specific approach to large scale implementation is required with an impact on cost. The provision of suitably high volume connection to the internet might also be an issue at periods of high usage.

## 2.3    Technical Review - Extensibility

Extensibility means the ability of the system to extend to a wide range languages, both for the material and the user interface, a large number of contributing organisations and a wide range of material types.

The system as it stands can support a limited range of languages used in the user interface (as demonstrated by Iberoamérica Digital pilot system) but the user interface will require adaptation to support a wider range of languages particularly where languages use different scripts. For example, different screen layouts for Chinese as opposed to Western European languages and the availability of virtual keyboards for the input of a range of scripts. This is not necessarily an easy technical development.

Similarly the system does not support material in different languages, for example, by providing the translation of metadata and search queries. Nor does it support language based processing, for example the recognition of language in the metadata or search query and the use of processing relevant to that specific language, such as separate indices and stemming algorithms. Again these developments are not easily achieved.

With respect to material types, the system is primarily focused on the supply of digitised photographic images through the use of thumbnails. The provision of other material types, such as digitised books or audio material, will require a more flexible approach to selecting and presenting summary material, otherwise the user will be faced with the prospect of examining each individual object to determine its relevance to their enquiry.

The overall conclusion must be that there is a substantial amount of development work required to create a service that extends across a wide range of language and material type.

## 2.4    Technical Review - Interoperability

Author: Bill Oldroyd
20th May 2010

From the point of view of metadata and data harvesting, the service is based on widely used interoperbility standards. Harvesting from web pages and OAI PMH are standard approaches widely available in library software.

The use of a simple metadata format based on Dublin Core is compatible with services such as TEL and Europeana. However, there are differences and complete rigid conformance would require the metadata to be converted. For example :

- the use of a language attribute to indicate the language of a specific element

- the use of local element names for metadata outside Dublin Core, such as links to the objects

NLG provides a bespoke web-service API for searching and data extraction. This API is easy to implement, but the fact that it is not using a standard API is unfortunate. It would be better if this API were made compatible with existing standards such as the SRU protocol or the OpenSearch protocol. For example the use of OpenSearch would allow the easy integration of an NLG search in the search box of nearly all standard web browsers.

## 2.5    Technical Review - Conclusion

NLG is a service based on widely used state-of-the-art software in a fairly basic implemenation. Scaling to support high levels of use is dependent on investment in sufficient hardware, though at high levels of data acquisition it might be stretching the capability of the index and search software.

NGL has limited scope for extensibility, for example to cover use in a range of languages. This is likely to require significant investment in the further development the software.

## 3.1    Service quality

Service quality in considered from a number of aspects:

- user interface – ease of use

- support for searching

- delivery of useful objects to the user

- reliability, service levels and response times

The NLG service provides a plain, straightforward and easy-to-use interface for simple searches of the material. More complex searches, including the specification of search fields and the use of Boolean operators, can be made through the search box, but how to do this is not explained.

The display of facets allows the user to narrow the search by the chosen fields, for example date, subject and institution.

The use of the "BACK" button allows a user to trace back through a search dialogue. The combination of these features allows a relatively simple but sophisticated search interface.

Author: Bill Oldroyd
20th May 2010

The record displays are simple and straightforward and obvious links lead the user to the source object. The ranking process usually provides a useful ordering of the records.

The interface of NLG is monolingual, but an implementation of the software for Iberoamérica Digital illustrates the use of the system interface in English, Spanish and Portuguese.

There is a variability in the access provided to the digital objects - some are usable (readable, viewable) and some not. It is a basic premise that users consult search services in order to consume the information or experience provided by the objects indexed by the search service. A major quality issue is therefore the extent to which a user, having searched NLG, can "consume" the objects found. The display of the objects may appear out-of-scope for the service objectives of NLG but it is very relevant to quality of service provided to users. There are many issues, for example the provison of digitised data with a resolution that makes the object readable, the manner in which users can browse the object, the language of the interface and so on.

Unlike World Digital Library, which addresses the issue of access to the digital objects, NGL refers the user to the local implementation of a specific object. Europeana and TEL do this as well. This creates problems in terms of the language of the interface, bad links that do not resolve (many Singapore links do not resolve correctly) and often an inability to browse the object sufficiently to gain any information from it. For example, the image of a large object such as a map is not displayed in a sufficiently high resolution to be able to read it.

Many of these issues are addressed by standards required by the World Digital Library or the Internet Archive Book Reader. This failure to address access to the objects as part of the service has been noted as a feature of user dissatisfaction in user studies for Europeana where a similar situation can arise. This will apply to NLG as well.

There is a second issue about the scale of the database. Most of the material has been collected as part of a collection process, which means in most cases that intellectual effort has been focused on a limited set of material in order to select and describe the items. In addition, in local systems providing access to a collection, there are often a range of special functions and descriptive information that supports the use of the material. This is usually in the language of the provider and implemented in a particular house style.

The current partners illustrate two different approaches to exposing digital objects through NLG.

Australia has exposed a large number of objects. These are not grouped into collections, yet clearly the material was assembled on the basis of specific collections. A lack of an overall description of the material, which might have been derived from a set of collection descriptions, leaves the user in the dark as to what is actually indexed. The only way of resolving this is to spend time exploring the database through search and browse, which takes time and leaves the user with the question as to whether there is nothing there or whether they are using the wrong terminology in their search.

The Library of Congress has provided records which briefly describe specific collections drawn from American Memory and other sources. This prevents detailed searching of the content of a collection directly from the portal but on the other hand gives the user an overview of the material included in the service and invites the user to search in the local interface. This creates problems in that this material cannot be included in a search using NLG and the interface may not be in the user language.

TEL collects and offers collection descriptions to the user. This is not a perfect interface, but

Author: Bill Oldroyd
20th May 2010

it gives the user and idea of the content (there are over 400 collections in TEL). Multilingual access is supported by translating all the scripts into the range of languages supported by TEL. This multilingual support is expensive in terms of translation cost and the complexity of the system. However, the sophistication of internet translation services is growing and these systems can in many cases provide the gist of the text being translated, sufficient for someone searching material. Translation can be applied to any web page.

If partners contributed collection descriptions which described the material included in the service it would greatly help users to understand the content of the service. However, the requirements for participation in NLG should be clear about the inclusion of metadata for the specific objects.

## 3.2    Service Quality - accessibility

There are two basic accessibillity issues:

- Pages do not re-render when they are scaled or character size is changed.

- Images do not have descriptive text alternatives associated with them. They have metadata alongside them, but is that the same?

## 4.    Benchmarks

The pilot system is implemented on a test system. It is impossible to perform meaningful benchmark tests in this situation. The best judgement we can make on the performance of the service is a comparison of the implementation with other services. As pointed out in the technical review, the design of NLG is similar to equivalent services. If a suitable hardware implementation is available the service will scale to provide a satisfactory response time.

## 5.    Operational Issues

Ongoing maintenance is not an easy issue to determine. Europeana is not yet an operational service and ongoing maintenance levels are not easy to determine. TEL is operational and requires maintenance support of 3 staff but these are involved in active development.

The data harvesting process is not automatic. Neither TEL nor Europeana has moved to a steady state or regular, automatic harvesting and this makes it difficult to predict the on-going requirement for staff, but at present TEL requires 2 staff associated with this activity.

NLG as presently configued requires little in the way of web-site management, but if the inteface becomes more sophisticated then this will become a more significant issue.

Coupling these roles with management, promotion activity and communication with partners, one might expect NLG to require a staffing level of at least 12 people.

## 6.    Rights Management

The service should ensure that organisations contributing material to the database can provide rights to use the metadata supplied and any "summary" information such as a thumbnail or other extract from a data stream. Rights management for the metadata and summary information is the responsibility of the contributing partners and the information should be made available through a Collective Commons licence.

Author: Bill Oldroyd
20th May 2010

As access to the objects themselves is controlled by the metadata providers, local access control will apply.

Introducing complexity in rights management is an issue that affects overall scalability of the service.

## 7.      Business Model

It is worth pointing out that if the scope of NLG was limited to providing search of collection descriptions provided by the National Libraries rather than a search of the digital object metadata, the scope and costs of NLG would be substantially reduced, probably by a factor of 100. This would not give users a unified search across the objects themselves, but it would make the user aware of the digitised material available from National Libraries, which can then be searched locally. There is little merit in merging the two types of metadata record in the same database.

However, assuming the overall objective of NGL is to provide a search of the object metadata there are a number of operational models that may be used :

- Central model

- Distributed model

- Other solutions based on

    o   Internet Search engines

    o   Library Service providers

There is a core issue as to whether making this data available in a set defined by National Libraries is a useful service compared with making the data available through a confederation of other sources — of course it is possible to do both.

It also raises the question whether the scope is too broad. Would smaller aggregations of material be more successful and useful in terms of a number of factors :

- Ownership

- Potential sources of funding and project support

- A community of interest, culture or language

The complexity of a world-wide solution addressing questions of language for searching and display also indicate a considerable challenge.

Some of these challenges are addressed by the major search engines and therefore it is appropriate to consider if using these services is not a better solution. For example, Google provides a range of translation services both for the search query and results page. Whilst these services may not be perfect, they will for most objects provide an adequate indication of the content. Furthermore, over time these services will improve.

The choice of this option requires consideration of the role of search engines with a world wide scope, for example, Google, Bing, Yandex, Baidu, Ayna and so forth.

Author: Bill Oldroyd
20th May 2010

This also raises the question as to whether information from National Libraries is better placed in systems alongside material from other archives, museums and other libraries, the greater volume of like material proving more attractive to user.

NLG should consider the relationship with established regional, subject and organisation services which have similar but more focused aims. A world-wide set of services could be coordinated through :

- Interoperability

- Focus of local issues for search, metadata etc

- Sharing access through federated search

It would involve the promotion of standards that would facilitate this interoperability.

The next steps would be:

- Identify potential partner organisations

- Set up cooperation to share, exploit and encourage the use of open source software

- Agree a mechanism for interoperability

**Conclusions**

This review has concluded that the technical infrastructure of National Libraries Global shares much in common with similar services, such as The European Library and Europeana.  There is certainly potential for cooperation between National Libraries Global and The European Library but, at the same time, the report points out that investment would be required to achieve scalability and extensibility.  Some initial thoughts have been given to the development of a business model and adequate resourcing but the recommendation is that this would need further work.

Author: Bill Oldroyd
20th May 2010