DATA PROCESSING AND SUSTAINABILITY WITH A LARGE-SCALE AGGREGATOR: THE UK ARCHIVES HUB

Jane Stevenson

Introduction

The Archives Hub[1] is a UK aggregator for archive descriptions, enabling researchers to search across descriptions from over 330 repositories[2], consisting of around 1,500,000 units of description. It takes in anything from brief collection descriptions through to complex multi-level item descriptions. It began in 1998 as a proof of concept and the official launch of the service was in 2001. Since then, the service has grown from representing around 120 higher education institutions to representing a wide variety of repositories, including specialist, local authority and business archives. Use has increased substantially over time, so that we often have over one million page views a month[3]. The team work closely with contributing repositories to support data creation and enhancement and to create a sense of community around the service. We aim to understand and meet researchers' needs, raising awareness and use of archives.

A New System

In 2014 we set out to replace our old system and interface. The Archives Hub had done well on the old Cheshire XML search engine[4] that we used since the service started, but it was becoming unfit for purpose, as technologies changed. We had over ten years of experience in developing the Archives Hub, and we had a very clear sense of what we needed going forwards. We were confident we could bring all our experience to bear, learning lessons from the past, building on strengths, to create something more robust, efficient and sustainable. This article is about how we set about the re-engineering of the Archives Hub.

Workflow

At the core of our thinking was the need for a new automated workflow. Up until that time, all the descriptions that we had taken into the Hub were manually processed. This may seem surprising, but when the Hub started in the late 1990s, things were very different, and the Web was still in its

[1] https://archiveshub.jisc.ac.uk
[2] https://archiveshub.jisc.ac.uk/search/?tab=locations
[3] Taken from Archives Hub logs, e.g. 1,047,994 page views in July 2018 excluding known bots
[4] http://cheshire3.org

infancy. Indeed, the team specifically included a 'Data Editor', whose job it was to check the descriptions as they came in, and, if felt necessary, make manual alterations, such as adding index terms and removing 'redundant' information.

Over the years, an unfortunate result of manual processing was to create major issues around version control and consistency. If you bring in descriptions and change them during ingest, but do not pass the changes back to the contributor, then you end up with different versions of the same thing. The level of manual work that we used to do was a useful way to interact on an individual level with contributors and to build up our knowledge of UK cataloguing practices, but it was clearly unsustainable if we wanted to take data in at scale. We required a workflow that gave us the necessary structure to process diverse data, with minimal manual intervention, whilst avoiding version control issues.

We did not have a clear and documented Hub content standard for the data that we ingested. In principle we adhered to ISAD(G)[5], and we had some additional recommended fields, but we had not properly enforced this. It would have been difficult to do this without an automated process, as checking individual descriptions manually is no substitute for the rigours of machine-based validation. As a result, we had many descriptions that lacked 'mandatory' fields, such as language and access conditions, and we had a number that used the same references, particularly at lower levels, which meant that only one would be visible in the user interface (the reference was used as a unique identifier). Whilst the Archives Hub had been very successful, and use continued to grow, we wanted to update the user interface, introduce new functionality, and think about future potential developments such as name authorities, that required consistent data standards.

What we needed was a system that gave us a balance between rigour and flexibility, automated processing and manual intervention. Our budget and timescale made it unlikely that something could be built from scratch to meet our quite exacting requirements. We set out a detailed specification that covered inputs and outputs, persistent identification, handling revisions, administration of data processing and all aspects of the web interface search, filtering and functionality. We also asked for solutions that could be applied across descriptions of archives, thematic content, repositories and names. Knowledge Integration (K-Int), a company based in Sheffield, with extensive experience of working with museum data, won the contract[6]. They already had a system that could be modified to suit our needs. Their CIIM (Collections Integration

[5] https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition
[6] https://www.k-int.com/

Information Middleware) was already used by a number of museums, and they were highly recommended by many of their customers. K-Int were confident they could modify the CIIM to allow for our particular needs. They also nominated Gooii as the front-end designer[7], and this worked very well, as both companies had already worked closely together on a range of websites and Gooii had experience with the cultural heritage sector.

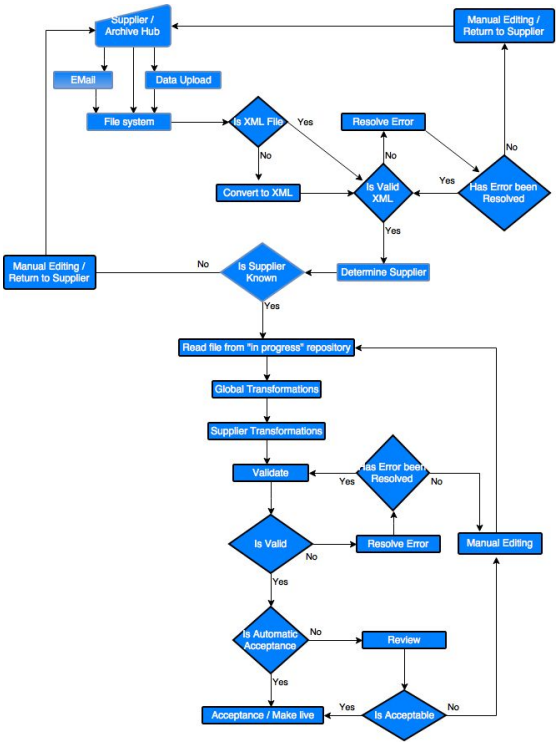

Fig 1: Archives Hub Workflow

The diagram (Fig. 1) shows how the CIIM implements a workflow that meets our needs. It is applied to all of our XML archival standards – EAD[8], EAC-CPF[9] and EAG[10]. We largely focussed on EAD during the initial phase of the work, as this is for the archival descriptions, which are at the heart of what we do. Our first job was to move 100,000 EAD descriptions (consisting of over 500,000 units) from our old system to the new one. The migration of the data gave us the opportunity to implement validation checks and undertake one-off normalisation work. So, this was not necessarily going to be part of the ongoing data transformation; some of the normalisation would be integrated into a continuing process and some would never need to be repeated.

Data Transformations

[7] https://gooii.com/
[8] https://archiveshub.jisc.ac.uk/ead/
[9] http://eac.staatsbibliothek-berlin.de/
[10] http://www.archivesportaleuropefoundation.eu/images/docs/EAG2012_TagLibrary.html#tl3c

As soon as the data comes into the CIIM, K-Int are responsible for implementing the basic checks: Is this XML? Is it well-formed? Can we identify the repository it comes from? If the description passes these checks then it reaches the 'transformations' section of the process, to apply more tailored processing that is specifically to fulfil Hub requirements. After these transforms are applied, the CIIM then checks whether the document is valid and can be accepted and made live, or whether it has errors or conflicts with another description, in which case the CIIM will apply a status to indicate this, and the description will not be published.

The transformations are the part of the process that the Archives Hub team control. They are XSLT files, each intended to fulfil a different purpose. The aim is to implement the Hub content standard. The data requirements we came up with were carefully considered, and took a good deal of time, effort and discussion to finalise. Our aim was to ensure that the descriptions worked effectively within the new Archives Hub user interface, but more than that, we wanted to provide API access and enable harvesting, so that we could fulfil a wider role of promoting the archives as widely as possible through different channels. Ensuring that descriptions work adequately within one interface is one thing; creating an interoperable set of data is quite another. In addition, it is only by clarity and rigour in processing that it is possible to work successfully with revised descriptions – to identify them as such and ensure that we do not take in duplicates. In the past we have had significant problems ensuring that revised descriptions replace older versions. Revisions may include changes to titles, references, and even filenames, as well as being significantly expanded so that they barely resemble the original description.

During this whole process we examined each of the most common fields that are used in archival descriptions and thought about what we required. We wanted to have a clear and justified rationale for all of our decisions, and we needed to be confident that we could implement them consistently. I shall just explain some issues relating to a couple of fields, as examples of the process that we went through.

Data Challenges

The unique identifier for each unit of description was the most challenging area, but the most vital for us to get right. We wanted something that would be persistent, and we debated for a long time whether the archival reference should be incorporated into the identifier, what the pattern of the identifier should be, and how we could ensure it was globally unique. We concluded that it was useful to have user friendly URIs that researchers could use to cite material, and therefore we did

incorporate the archival reference. We also used a pattern that we felt was in line with the recommendations for 'cool URIs', thinking in terms of 'simplicity, stability and manageability' as well as the future potential for Linked Data approaches.[11] In retrospect, the incorporation of references has given us a number of problems; an opaque identifier would have been a great deal easier to implement, and yet, it does not seem well suited for referencing materials in publications and elsewhere. In addition, we wanted to give our contributors access to Google Analytics statistics, and by using archival references they can identify collections much more readily. For example, the URI https://archiveshub.jisc.ac.uk/data/gb248-gua works well in citations, in statistics, and in other places. The contributor is likely to know which collection 'GUA' represents; whereas an opaque identifier looks far more random, e.g. https://archiveshub.jisc.ac.uk/data/5423a1c4-bcfe-3c21-8dcd-0ebf353a9207.

We had problems with duplicate references due to cataloguing errors, and a particularly tricky problem with duplicate reference such as *GB1867ab* and *GB1867ab* where one is reference 'ab' from repository '1867' and the other is reference '7ab' from repository '186'. We decided to introduce a hyphen between the repository code and reference, to create GB1867-ab and GB186-7ab. There were tricky characters such as slashes, ampersands and asterisks to deal with, and references that were over 100 characters long (which do function, but they are far from ideal as user-friendly URIs).

Cool URIs don't change, and the main problem with using the reference is that the contributor might decide to change it, despite the repercussions of doing this. Unfortunately, the situation was further complicated by the fact that the Hub team used to change all the file names on ingest to a simple consecutive numbering, so we could not use our existing filenames to help us identify revisions. In the end we had to implement a means to compare files by a combination of file name, reference and title. This works well, so that in the majority of cases revisions are identified and there are no clashes. The CIIM warns if a title has changed, for example, and gives you the option to confirm the revision, create a new additional entry, or reject the update. But some revisions may still not be identified, if so much has changed as to make the comparison with a current file impossible.

Another problematic area was the lack of consistent divisions, or lack of any divisions at all, between entries for things like language, creator and index terms. We had to analyse the data and aim to create transforms that would split terms with dividers such as commas, semi-colons, brackets and slashes into separate fields. Library of Congress Subject Headings add an extra complication because they often include dividers within a subject entry, such as 'Biscuit industry, Great Britain'. Having

[11] W3C Cool URIs for the Semantic Web, https://www.w3.org/TR/cooluris/#cooluris

more than one subject in a field might show up as 'Biscuit industry, Great Britain, Cheese industry, Great Britain' or as 'Biscuit Industry – Great Britain, Cheese Industry – Great Britain'. The latter is easier to divide. It had to be done on the basis of analysing the patterns used by each institution; we cannot apply solutions like this across all of the data, and even then, we cannot hope to correct all the data. Currently, we are still working to find a suitable fix for issues with different amounts of white space within index terms. The aim of all this complicated work is to provide a better facility for a researcher who wants to search by subject, language, creator name, etc., and to give us the potential for more ambitious interface developments. It is an ongoing process of analysis and improvement.

Data Standards

A description is not accepted if it does not have all mandatory elements. We apply ISAD(G) mandatory fields, apart from name of creator[12]. We made the decision to drop name of creator as mandatory because thousands of descriptions on the Hub did not have a creator name, and it is not always possible to identify a creator. There were also hundreds of descriptions already on the Hub that did not have language, extent, access conditions or scope & content, but we decided that we would aim to populate these at the top level of description (collection level), to give researchers a more consistent experience and help them to identify relevant materials for their research. This meant asking a number of contributors to revise descriptions that may have been on the Hub for many years. Therefore, it was important for us to justify this approach; to explain exactly what we were doing and why. The benefits of bringing all the data up to a certain standard needed to be clear. In fact, we did not have any dissenters; it was simply a case of contributors needing to find the time, amongst all their other priorities, to revise their descriptions. For many of them, making improvements was seen as a positive thing. In addition, a number of contributors were happy to replace all their old descriptions with new ones, thus by-passing the problems of missing data, and at the same time giving us up-to-date content.

This whole process was complex, and it was made a great deal more difficult by the fact that we had to implement the new system at the same time as we were making decisions about the data and also completely re-writing our cataloguing tool. I would thoroughly recommend making decisions on data standards first of all, and not at the same time as a migration and system update. Several times we had to adjust what we were doing in terms of data processing because we decided to alter our data standards, in order to ensure that we were being practical and creating something that was

[12] https://archiveshub.jisc.ac.uk/isadg/

sustainable. For example, we only decided to drop creator as mandatory after a few months of work; it would have been less than ideal if we had already insisted on some contributors populating creator name in order to keep their data in the Archives Hub. At the end of the process there are still some requirements that we could potentially change. Controlled level values are a good example of a mandatory requirement that arguably don't provide enough benefit to justify the corrections that are required to the data. However, addressing level values did allow us to implement a filter on the Hub interface to search by collection only, item only, or 'sections' in-between the two.

Data Pipelines

A selection of the XSLT transformations we apply to the data are put together make up a 'pipeline' for each contributor. The pipelines that we create fit into three categories. We have global pipelines, which consist of transforms that are applied to all the data that is uploaded. We have group pipelines, which can be applied to a group of data, such as data exported from a specific archive system, and we can also implement individual pipelines, for just one contributor.

The global pipelines include things like checking that normalised dates are properly formatted, changing a hyphen divider to a slash divider, consistent with ISO 8601[13], checking language codes are correct and changing level values to be consistent (e.g. 'Sub-series', 'Sub-Series', 'sub series', 'SubSeries' all become 'subseries').

The group pipelines are used for exports from archival management software. At the moment we have two groups for exports from the Axiell Calm system[14] - a default one, and one for Calm descriptions from welsh repositories, which includes some additional transforms relating to language. We also have a configuration file that we use for Calm descriptions that allows us to select options relating to the treatment of Calm references. Calm has automated references, which use slashes to indicate hierarchy, and also 'alternative references' that are free text. Institutions may use just one of these as the display reference, or a combination of both. Our configuration file allows us to pick from eight options in terms of using the reference for the URI and for display within the interface, e.g. use only the Calm reference, use only the alternative reference, use the alternative reference if there is one, and if not use the Calm reference.

[13] https://en.wikipedia.org/wiki/ISO_8601
[14] https://alm.axiell.com/collections-management-solutions/technology/calm-archive/

We try to avoid using individual transformations, as we aim to have groups of contributors in each pipeline, which cuts down on the administrative overhead, but sometimes it is necessary, to address repository-specific data issues. A contributor such as the British Library is quite bespoke, so they need a transformation that is just for their data.

Community Engagement

It has been really important to engage our contributors and get them onside with this whole process. The Hub team have always sought to build a sense of community amongst our contributors, encouraging repositories to see themselves as stakeholders in the service. This has proved to be very beneficial with running a successful service, creating trust and managing expectations. Our plans and proposals were well received, and many contributors appreciated the data analysis that we carried out for them. It is hard for many of them to do this kind of thing within commercial systems; that is, to check which fields are used, where there are data errors or issues, and where there are inconsistencies. Furthermore, many repositories just do not have the expertise to run analyses on their data. Whilst we were asking them to put time and effort into make changes, we always provided a clear specification of exactly what we need, and overall the process helps people to improve their data.

We intend to provide access to our administrative interface (CIIM)[15] for all of our contributors, and this is another way for us to engage them in what the Hub is doing, giving them more control to view and search their data, and enabling them to instantly un-publish a description if they wish to. A number of contributors are in a position to upload content themselves and use the administrative interface for data analysis; others will continue to use our in-house cataloguing tool and their descriptions will be uploaded to the CIIM automatically when they want to make them live. It is vital that we don't expect all contributors to have technical expertise, or to have the time to administer their data on the Hub, but at the same time we want to provide this option for those who see it as a benefit.

Reuse and Data Potential

One of the key motivators for this whole approach to data processing was to ensure that we could provide content through our API and through OAI-PMH harvesting. We had already agreed to

[15] https://archiveshub.jisc.ac.uk/ciim/

provide all of our content to the Archives Portal Europe (APE)[16], so we had a concrete use case. As the Archives Hub is a national aggregator, and as we work with EAD, we were in a good position to contribute to APE. It was a good opportunity to see how re-usable our data really was. APE has certain requirements that we had to meet, but this required relatively little work compared to each institution trying to contribute their data themselves, and once it was set up, it became an automated process, with harvests taking place each month. It was gratifying to realize that we were in a better position with our UK repository data than most European countries, and that our aggregation is probably the best developed in Europe in terms of data ingest, so that we now contribute data to APE from significantly more repositories than almost all other countries.

Our experience of running a national aggregator had taught us that it is important to think about the potential of the data. This is a somewhat nebulous concept, and yet it is worth being aware that decisions made during the initial data processing will have implications for what we can do down the line. For example, we thought about the kinds of visualisations we might be able to implement. We wanted a map showing the location of collections. We were adding latitude and longitude to the information about repositories, so we knew this was do-able. We also considered a graphic representation of collection strengths. For example, where are the repositories with higher concentrations of archives about 'Victorian theatre' or 'urban development '? For this to work, we needed to use the 'extent' field. We proposed to analyse the extent information and find ways to make it more consistent. However, it became apparent that the variations in the way extent is described made this impossible. We could deal with reasonably consistent use of various units - boxes, folder, cubic metres, linear metres, files, items - in a way that would represent them in terms of relative size, albeit in a fairly rough way. But extent often includes things like 'a page of notes', '4 pieces', '54 sheets', '500,000 records', and often mixes genre and dimensions with size in an unstructured way. This is an example of the drawbacks associated with a more descriptive way of describing materials. It works well for the human eye, but it mitigates against any kind of useful machine-based processing.

Conclusions

It is hard to get archival systems to modify their structure and output for our benefit, and it is hard to change cataloguing practices. We had to come up with an approach that recognised this. We could not ask contributors to make all the changes that we wanted; we had to take on the lion's share of the normalisation work and only ask them to make changes that we could not or should not make. If

[16] https://www.archivesportaleurope.net

a mandatory field is missing they usually need to populate it, but for something like access conditions we can create boilerplate text for them on ingest, or if language is missing we can add 'English' as a default, if appropriate for their descriptions. In many ways, our project was very ambitious, as we were trying to deal with standardising legacy data from many different systems, and also data which had been processed by the Hub over time, without consistent standards. Also, the idea of 'standards' 15 years ago was very different from now, with the emphasis on web-friendly re-usable data and inter-connectivity.

We are happy that we have very largely achieved what we set out to do. The Archives Hub interface now has effective simple and advanced searching and far more ability to filter search results. We wanted the researcher to be able to undertake a broad search, and then narrow it down by time, subject, creator or repository. On the old Archives Hub we never managed to get a date search or filter to work successfully, but our work to standardise dates has addressed this problem. We also have a filter by digital content, and we are going to look at how we can further develop search and display of digital content. We have tabs for different types of search, including archive descriptions, themed collection descriptions and repositories, and we can extend this to names and potentially to other entities. The repository descriptions all include latitude and longitude, so we have a map showing the location of search results. Every item has a persistent identifier and we have recently introduced a citation facilitation that takes advantage of our persistent URIs for citing Hub pages.

As a result of the automated workflow we have taken in descriptions from almost all Welsh repositories, mainly exports from the Axiell Calm system. We are also working on exports from other systems, such as AdLib[17] and Artefactual's AtoM[18]. Once we have analysed data from a number of repositories and set up appropriate transformations, we should be able to take data from all users of a system (unless they use it in an unorthodox way). We know we can easily tweak the pipelines if necessary, such as adding processing for a new data field or dealing with a novel pattern that we have not seen before.

As well as providing all descriptions to the Archives Portal Europe, we will be providing data to a Welsh archives portal, which is being developed, we are working with Scotland on a pilot project for data ingest, and we are looking at providing descriptions to other portals, such as the European Holocaust Research Institute. We can also create local interfaces for contributors using the Hub API[19]. So, our aim of using the Archives Hub descriptions for our own interface and for different purposes

[17] https://alm.axiell.com/collections-management-solutions/technology/adlib/
[18] https://www.artefactual.com/services/atom-2/
[19] https://archiveshub.jisc.ac.uk/microsites/

is realised, and we believe that we have fulfilled our intention of allowing for future developments and taking opportunities for connecting data in different ways.