

Workshop Aim

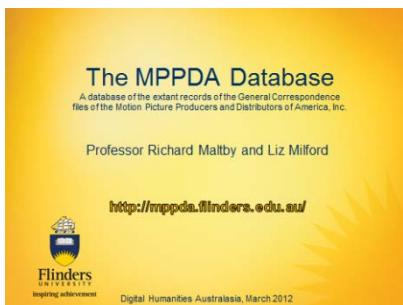
This workshop will use the development of a data management plan as the basis for looking at the different issues which have to be considered in the context of data management. Creating a data management plan provides a means of bringing a systematic approach to managing your data and ensuring its longevity and potential reuse. Such a plan typically includes: thinking through all of these issues before starting the research project can save you time and effort along the way and ensure that the data is managed both efficiently and effectively for the duration of the project and into the future.

MPPDA Presentation

Showcasing how data management supports re-use and curation.

The newly developed ANDS SC22B project funded MPPDA database now makes the material acquired by Professor Maltby available in an enhanced form, providing equitable but secure access to this valuable resource and a vastly improved way to browse and search for specific MPPDA records. This session will outline the project, how data management played a pivotal role in the project outcomes and the benefits which have resulted.

[Slide 1: Opening screen]



Apart from Ruth and me, probably no-one in the room has heard of the MPPDA (Motion Picture Producers and Distributors of America, Inc.).

[Slide 2: Hays]



MPAA – the copyright police – Valenti

Precursor organisation, best known for the Hays Code – self-regulation not censorship –

Much bad history written about this.

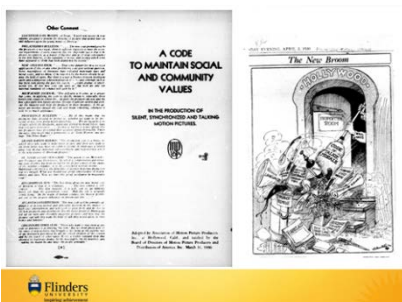
The functions of the MPPDA, and Hays' role as spokesman for "the organised industry," remain as poorly understood by most film historians as they were at the time, oscillating as they do between a vision of Hays as "the Czar of all the Rushes," with dictatorial power over the industry, and an equally spurious vision of Hays and the Association as ineptly outmanoeuvred by conspiratorial Catholics in the carefully choreographed public relations performances that took place in the summer of 1934.

Actually, the MPPDA managed the industry's politics.

So 28 years ago, I started a research project on the Hays Code, with a Fulbright scholarship to research in the US. Eventually this led me to the MPAA in New York, where I discovered ...

the General Correspondence files of the Motion Picture Producers and Distributors of America, Inc., covering the period from 1922 to 1939.

[Slide 3: MPPDA samples]



In 1965, the Motion Picture Association of America copied the General Correspondence files onto microfilm, and disposed of the original documents. The archive was then stored in the MPAA's New York offices, where it was largely unavailable to researchers.

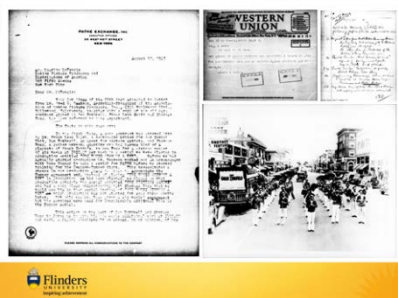
[Slide 4: MPPDA samples]



Other than the headings used on the folder in which the original material was stored, no inventory of the contents was made. The recording material used was 16 mm. negative microfilm, lacking

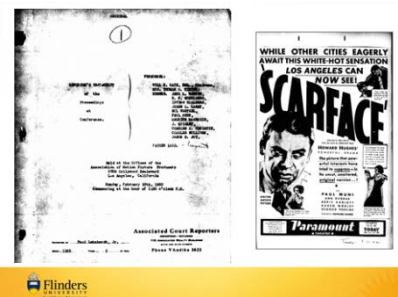
frame or edge numbers, making it difficult to find individual documents on the microfilm even if the researcher knows them to be there.

[Slide 5: MPPDA samples]



16 mm. negative microfilm – the most horrible data to access information from you can imagine. Quality is horrible – mostly fourth- or fifth-skin carbon copies, but legible to humans.

[Slide 6: MPPDA samples]

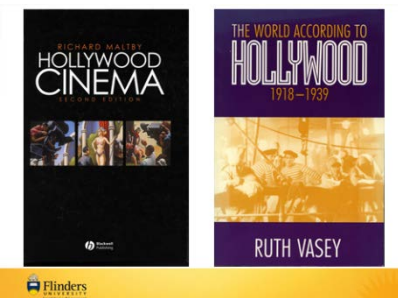


Why this stuff is exciting:

Although the documents in the MPPDA's General Correspondence files represent only a small fraction of the Association's output, they nevertheless represent an immensely rich source of information about the political history of the industry and a number of its key activities and decisions. They include extensive correspondence and other documentation relating to industry policy, public relations, trade practices, external censorship and self- regulation. The great majority of this material is unavailable from other sources.

[Slide 7: Publications]

Further outcomes in literary form



[Slide 8: Technological change....]

Building of the database over the years.....



[Slide 9: Early databases]

Building a database began as a catalogue, simply to make the material accessible for me (and Ruth and me) – no edge numbers, so constructed by document or collection of documents, rather than by individual frames



1984 Apricot – Cardbox – and Xeroxes. – Cardbox was a flat file database, but it indexed everything, so you didn't know how to use it, since everything was a keyword.

In 1997, when we moved to Flinders, we began a project to digitise the archive. It was just about possible then within a \$10,000 grant to buy an affordable scanner, scan each frame and Photoshop them into positives – and, of course, to build a more comprehensive database that identified individual frames. 35,000 frames.

Cardbox was now dead, so we rebuilt the database in Filemaker Pro.

This became an ARC project, with research assistance to complete the database. We've ended up with 3,500 records. We've continued to have fantasies of the documents being readable by OCR, but they're not.

The catalogue comprises about 4,000 records, with each record describing a document or cluster of documents. The document clusters reflect a combination of the original filing organisation of the material and our editorial decisions about the relationship between material and its inherent research interest. Some records, therefore, refer to individual pages, while in a

small number of cases, a record can identify a 200-page transcript of a meeting or 300 pages of material on import tariffs in Argentina in 1939.

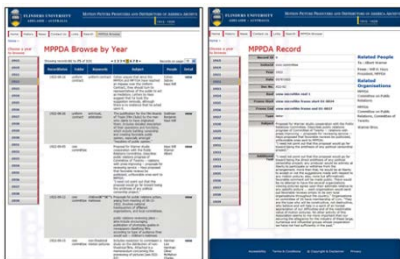
Large quantity of transcribed documents

In the later stages of this project, we brought Tim in to build us a proper relational database in Filemaker. However, since at least 2000, we had the idea of making the database available on the web.

Issues: the copyright police;

The need to make images available, since not everything is transcribed.

[Slide 10: First Web version]



In 2007-8, with Tim, we developed a prototype Web-based version, recompiled in MYSQL (version 5) with PHP (Version 5) and a simple Content Management System to generate the display. This archive was to be accessible via a “Search” page and a “Browse” page. The Search function was planned to allow full-text searching. The Browse page layout provided a summary description of the document, together with information on the document’s date, its original location, the people mentioned in it, and the keywords by which we have indexed it.

From this summary, users were able to click through to a more detailed record, which described the document in more detail, listing any films mentions, identifying the type of document and its author and recipient, and in some cases including editorial comment and a partial transcription.

- This early web based database was the basis for the current MPPDA Archive of today. This can be found at <http://mppda.flinders.edu.au/>.

[Slide 11: MPPDA Archive Website]



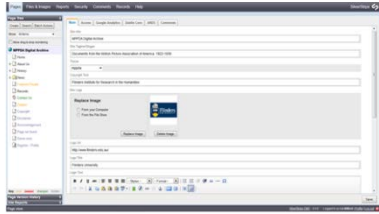
- The new site not only contains direct access to the MPPDA archive documents via the Records tab but also biographical information for people and organisations, news and historical information.
- Records in this dataset have been transformed to a format that is machine-readable and standards-compliant to support scholarly use and also make the complete MPPDA dataset, including digital images of archival material, available to authorised academic clients.
- Special interest areas relating to Hollywood industry policy, public relations, trade practices, external censorship and self-regulation can be readily accessed via the MPPDA data. This will, in turn, lead to increased citation of the MPPDA data sets, and the publications associated with those datasets, and enhanced access by Flinders researchers to research data in Screen Studies.

Some technical information (if needed):

- The web site as a minimum conforms to the Web Content Accessibility Guidelines 1.0. Standard HTML metadata markup is included in each page and supplemented with Dublin Core metadata tags and values (defined by Flinders University in line with metadata markup policies).
- The SilverStripe CMS is designed to work well with Firefox 3.0+ and Internet Explorer 7.0+ and provides satisfactory experiences in Apple Safari and Google Chrome. It also works well across Windows, Linux, and Mac operating systems.
- The MPPDA web application meets W3C guidelines for HTML, CSS, JavaScript, and accessibility, in addition to meeting specific guidelines, such as e-government requirements.
- The HTML based output of the web site conforms to XHTML 1.0 Strict which is the XML equivalent to strict HTML 4.01, and includes elements and attributes that have not been marked deprecated in the HTML 4.01 specification.

- The CSS used to display and style the xhtml conforms to Cascading Style Sheets Level 2 Revision 1 (CSS 2.1) Specification.

[Slide 12: MPPDA Back end]



- Authenticated project administrators and software engineers can access the MPPDA back end at <http://mppda.flinders.edu.au/Security/login>.
- This has been developed using a Silverstripe CMS to value-add to the project by allowing administrative users to administer html pages and forms and provides the underlying framework for the MPPDA data model.
- Our Three Fin Media consultant, Tim Cavanagh, has been instrumental in this process and has worked closely with Silverstripe to ensure this has been tailored to our needs.

Technical Information (if needed)

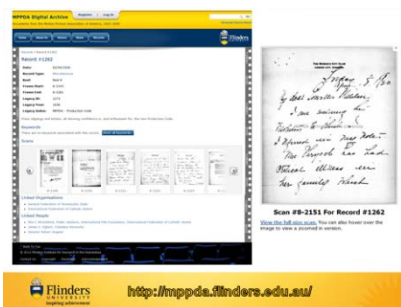
- The SilverStripe software is licensed under the open source (revised) BSD license. This is based on Model-View Controller (MVC) which is a software architecture which separates the data (model), the display or UI (view) and the business rules (controller).
- SilverStripe was chosen as the PHP framework for this project as it is built on the backbone of an object-oriented PHP5 web framework called Sapphire.
- The MPPDA data model is furthermore extracted to use Object-Relational Mapping (ORM) and this allows the project to be database agnostic and can therefore use a range of applications such as mysql, MS SQL Server, Oracle etc. The use of an ORM allows the data model to be exposed via a variety of protocols and Application Programming Interfaces (API). Elements of the project metadata are serialised to an xml file to support the RIF-CS DTD.

[Slide 13: Record example]



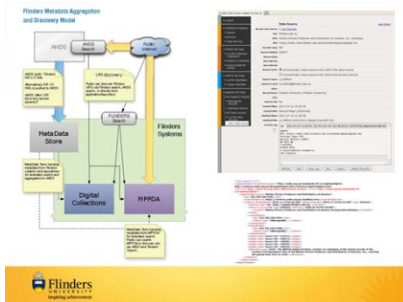
- In this new version of the MPPDA database, authorised users can now click through to a thumbnail of the digital image of the document, which provides enough information to decide on whether to download the full-size image.
- The process to authenticate access to the MPPDA database has been discussed at length with key personnel from the Flinders University Library Special Collections.
- Access to scans and transcripts of the archive documents described in the database is restricted to bona fide researchers, at the discretion of the Flinders University Special Collections Librarian. The database website directs users to a form on which they may apply for permission to access the material, which is password-protected.
- Acknowledgement of the Flinders University Library Special Collections should be made in all publications using material from any of the manuscript collections and the Library reserves the right to keep records of usage of the collection.

[Slide 14: Record example]



In a significant number of cases, users are also able to click through to a Word or Pdf file containing a transcription of some or all of the documents in the record. Full-size images are, on average, 450 kilobytes, and the whole archive is about 18 Gigabytes.

[Slide 15: ANDS Project models, RIF-CS record, metadata]



- RIF-CS record creation to make the archive readily discoverable
- The published RIF-CS metadata (including Collection, Party, Activity and Service metadata) for the MPPDA dataset has been created on the ANDS online Services Sandbox and published to Research Data Australia. This MPPDA metadata record is being moved to the Flinders University research data metadata repository (ReDBoX) and both editing and harvesting to the ANDS registry is possible from this system.
- Integration with Flinders University systems and policy is closely linked by the ANDS administered Seeding the Commons and Data Capture Projects. The clear need and overall drive for improved management of research data at Flinders University has been emphasised by these projects and the setup of a data repository and a metadata store and RIF-CS records are integral parts of this process.
- A key technical outcome for the MPPDA project includes the integration of the MPPDA specific data model into a modern object orientated web application framework that can easily be extended. In this instance an open source programming language and framework (PHP - SilverStripe) and database (mysql) has been implemented, enabling potential re-use or cloning of this system in future research projects.

[Slide 16: Acknowledgements]



Primary learning on how data management has played a pivotal role in the project outcomes

1. About the organisation of the data itself. How to present it.

2. Learning about what is required with research data management policy and servers. Integration and coordination of university-wide research data management policy work with other project work specific to the MPPDA database.
3. Need to find adequate database and file space from the outset as we struck some internal university problems relating to space and finding a suitable server to play on in development modes. In the end this had to be self funded as ANDS Seeding the Commons Funding did not cover this aspect and the lag in potential funding to set up repository space at Flinders did not correspond with MPPDA project phase completion.
4. Universities are struggling with the need for storage for a variety of e-research outputs. In fact, Universities are struggling with what storage they are going to provide full stop. Add in the complexity of a metadata store or performance requirements, at a time when most Unis are pushing large storage off to third parties such as Google, and you have a very difficult problem.
5. We were fortunate in having an established relationship with a database developer who also had a working relationship with Silverstripe – but we outsourced the development expertise and thus didn't acquire the corporate knowledge we might have done working in a different way.

[Slide 17: Credits and Contacts]



Benefits that have resulted from the most recent version of the database

1. A number of institutional benefits and implications can be noted as a result of the MPPDA project. These included an increased understanding of policies, processes and systems, training around management of research data at Flinders University, increased recognition of Flinders University as the agent for the preservation and curation work undertaken to make the MPPDA dataset available and consistent (and persistent!) moves to upgrade technical infrastructure to facilitate storage of Flinders University research data and creation and storage of metadata describing Flinders University research data. These fruitful collaborations with other university departments, personnel and projects have opened up

discussions for both future funding applications such as ReDS and the set up of an eResearch centre at Flinders University

2. The wider academic and humanities research community now have the opportunity to discover and re-use this fine MPPDA data resource via Research Data Australia with specific key research areas to directly benefit from the project including a number of ANZRC coded topics.
3. Special interest areas relating to Hollywood industry policy, public relations, trade practices, external censorship and self-regulation can also be readily accessed via the MPPDA data. This will, in turn, lead to increased citation of the MPPDA data sets, and the publications associated with those datasets, and enhanced access by Flinders researchers to research data in Screen Studies.

[Slide 18: Close]

