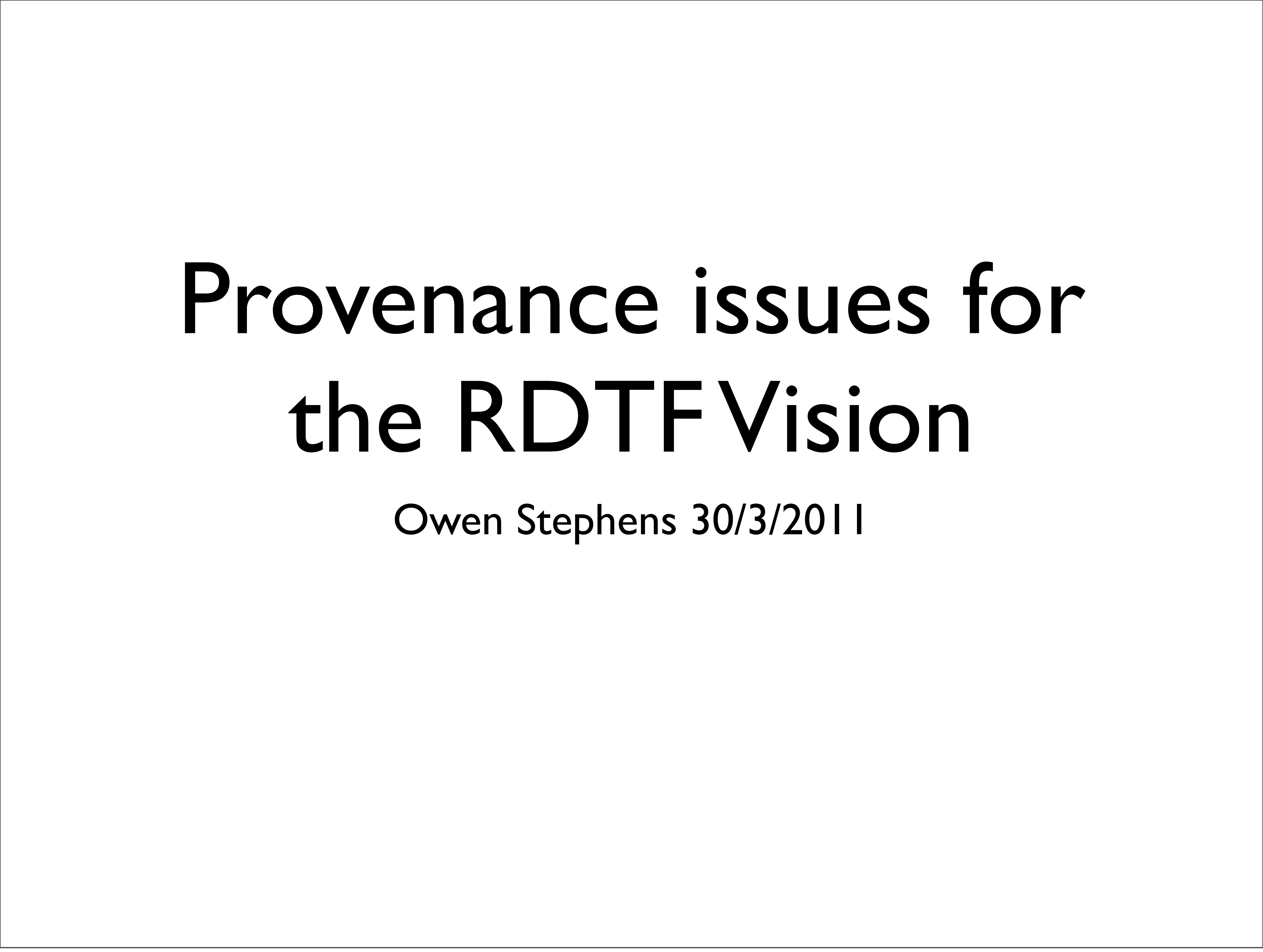


Provenance issues for the RDTF Vision

Owen Stephens 30/3/2011

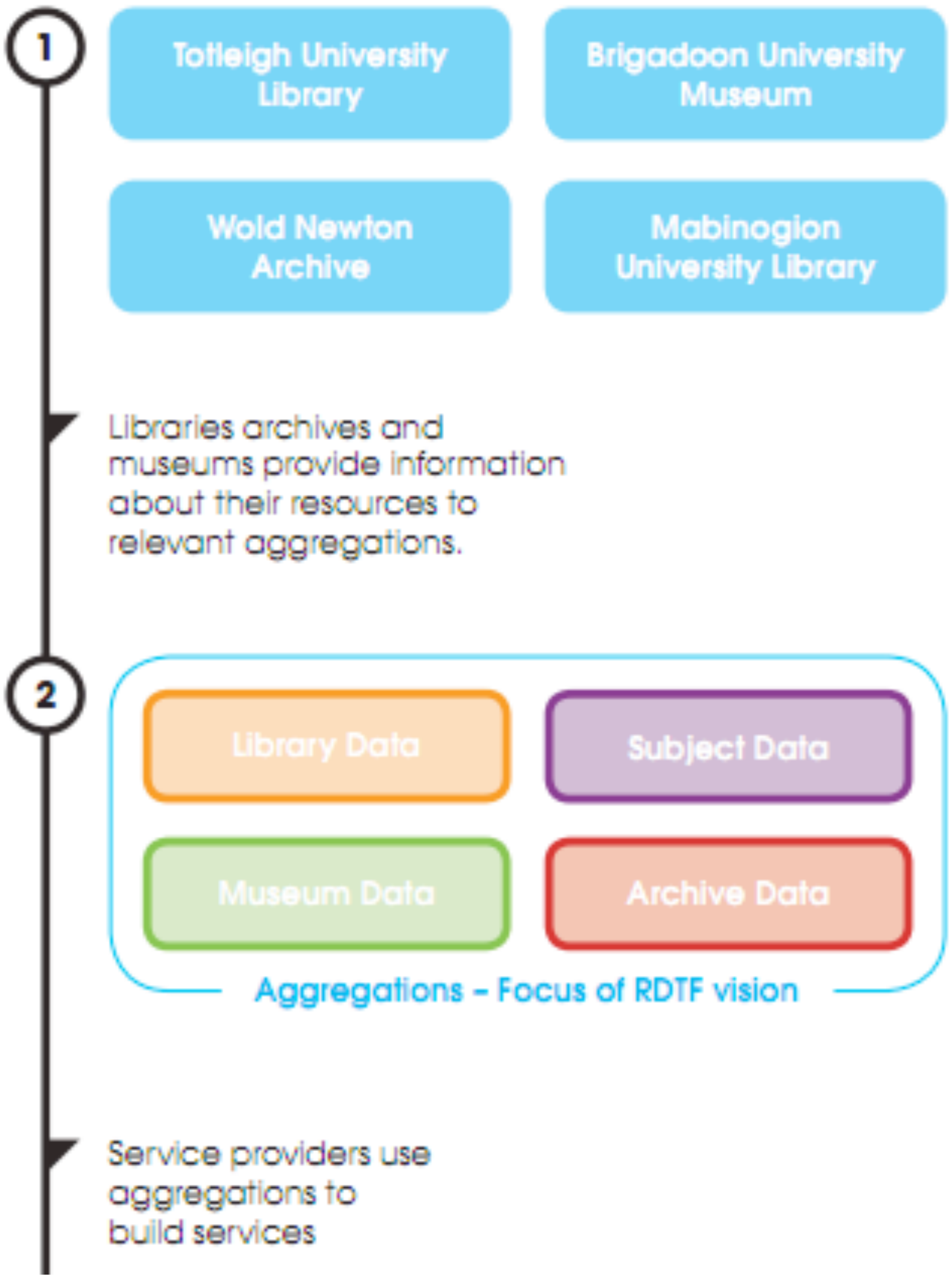
Wednesday, 30 March 2011

Briefly about me:
Librarian/IT
Lucero project with Mathieu D'Aquin + Fouad Zablith / KMi

What is RDTF?

Wednesday, 30 March 2011

The ‘Resource Discovery Task Force’ – set up by JISC and RLUK (Research Libraries UK) to “discuss what needs to be provided to help people discover and access items from Higher Education Libraries, Museums and Archives throughout the UK.“ (http://rdtf.jiscinvolve.org/wp/2009/06/26/hello-world-2/)



1

Totleigh University Library

Brigadoon University Museum

Wold Newton Archive

Mabinogion University Library

Libraries archives and museums provide information about their resources to relevant aggregations.

2

Library Data

Subject Data

Museum Data

Archive Data

Aggregations – Focus of RDTF vision

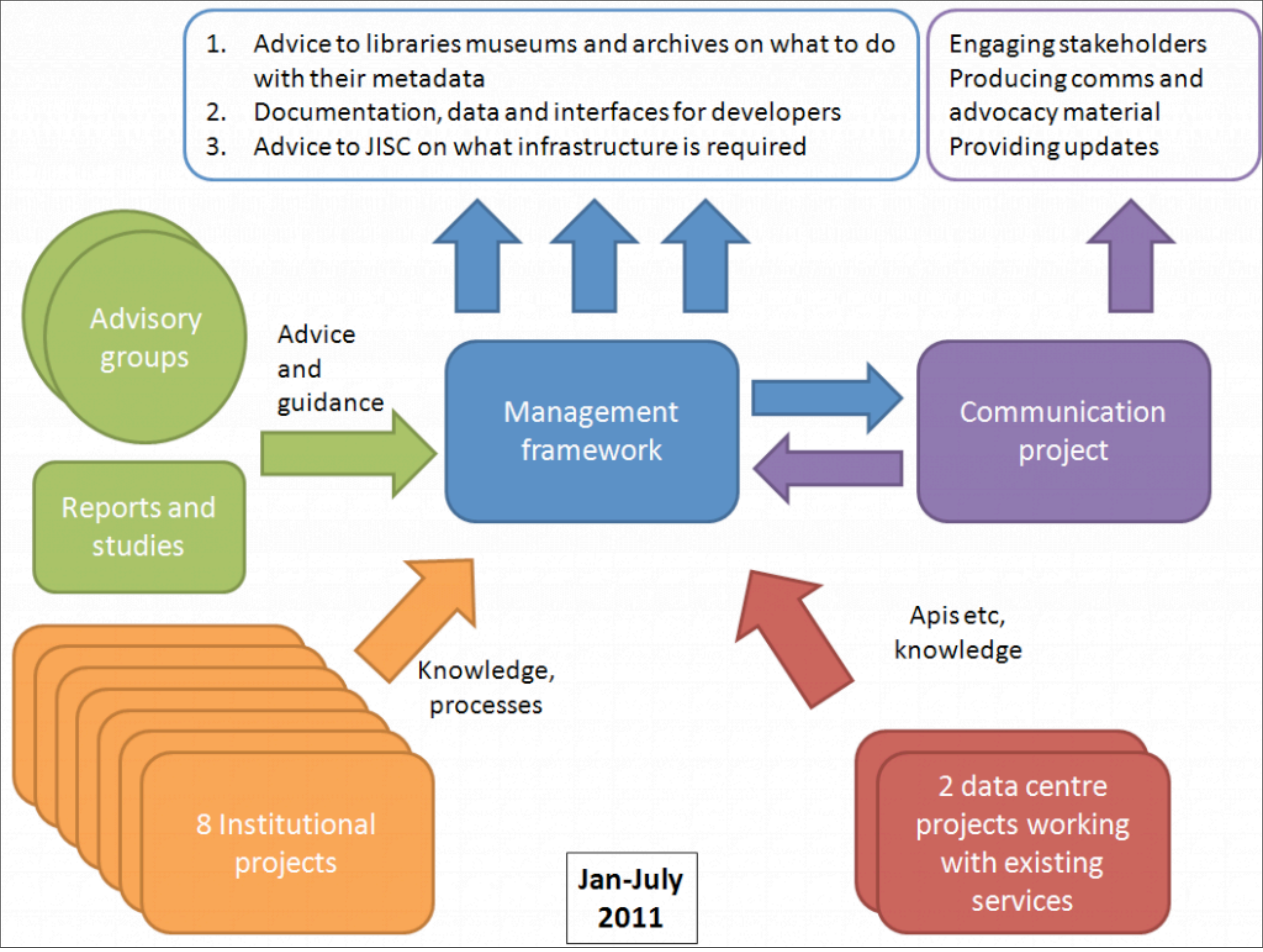
Service providers use aggregations to build services

VISION: UK RESEARCHERS AND STUDENTS WILL HAVE EASY, FLEXIBLE AND ONGOING ACCESS TO CONTENT AND SERVICES THROUGH A COLLABORATIVE, AGGREGATED AND INTEGRATED RESOURCE DISCOVERY AND DELIVERY FRAMEWORK WHICH IS COMPREHENSIVE, OPEN AND SUSTAINABLE.

The diagram to the left illustrates the vision and highlights that the focus is aggregations of content.

Wednesday, 30 March 2011

Result was a 'Vision' document, now leading funded work to try to realise this vision.


1. Advice to libraries museums and archives on what to do with their metadata
2. Documentation, data and interfaces for developers
3. Advice to JISC on what infrastructure is required
Engaging stakeholders
Producing comms and advocacy material
Providing updates
Advisory groups
Advice and guidance
Management framework
Communication project
Reports and studies
Knowledge, processes
Apis etc, knowledge
8 Institutional projects
Jan-July 2011
2 data centre projects working with existing services

Wednesday, 30 March 2011

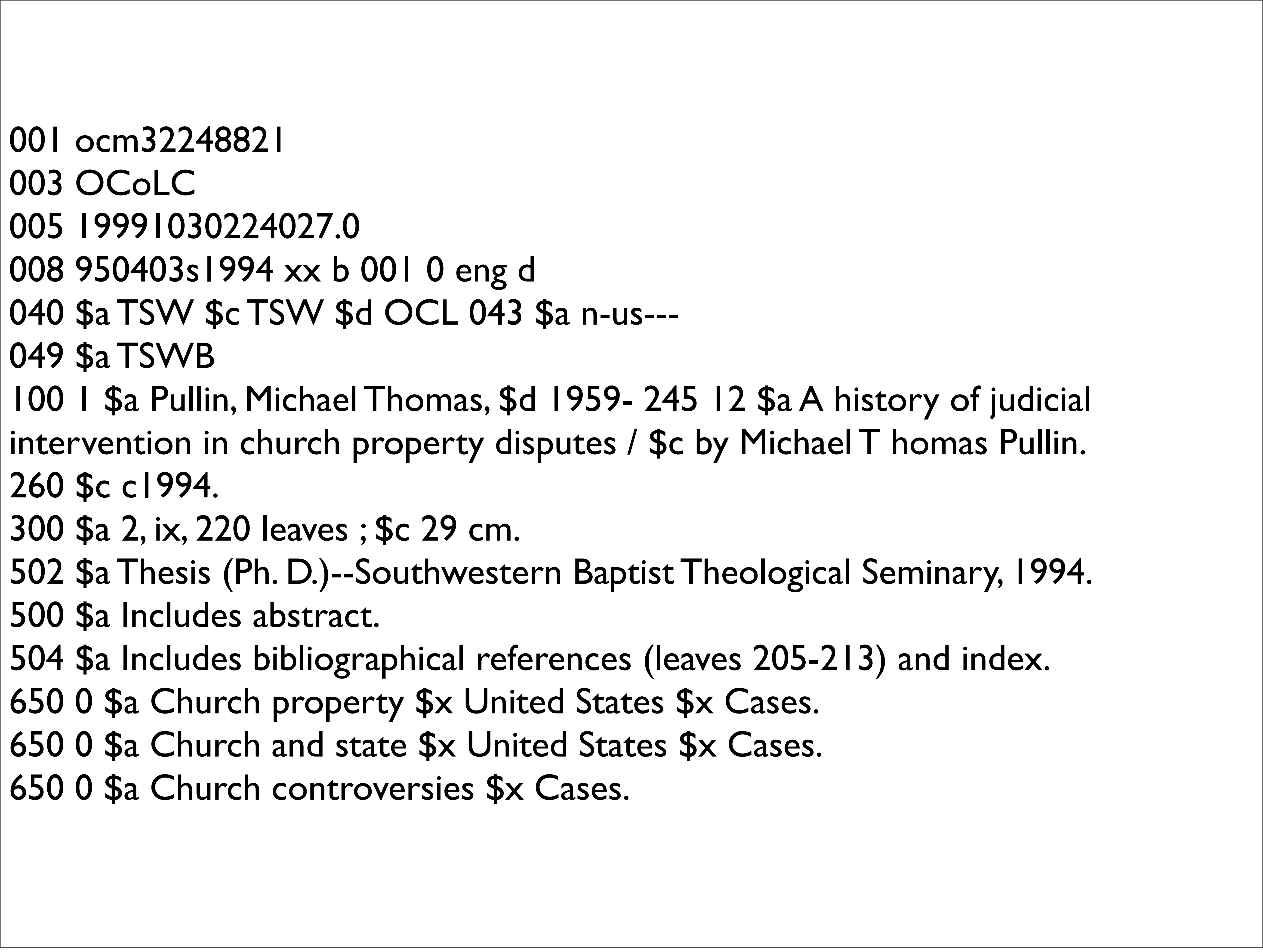
First stage – funding for ‘management framework’ project; communication project; and 8 projects going to get data out there

8 Institutional Projects - at least 5 intending to used Linked Data...

What kind of data?

“Why are you using Linked Open Data?”
“What are the characteristics of your data?“

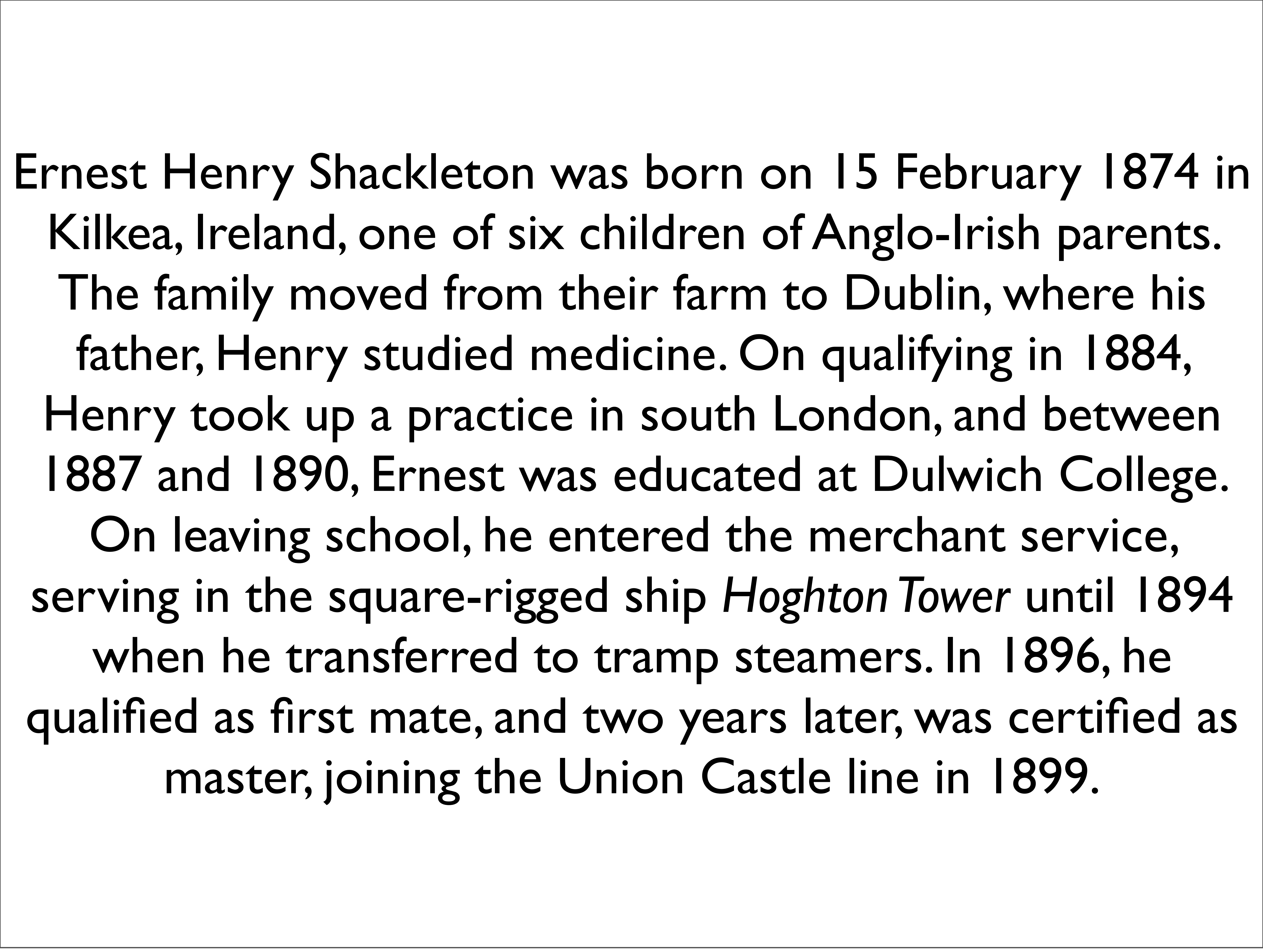
Variety, but bibliographic; archival descriptions; art/museum – METADATA (generally, at the moment)
Summary available at http://rdtf.mimas.ac.uk/newsletter/rdtfnewsletter01-march2011.pdf
NOT EXCLUSIVELY LINKED DATA
AIM25 Open Metadata Pathfinder - King’s College London
Demonstrating the effectiveness of opening up archival catalogues for automated linking and discovery.
Comet - Cambridge University
Releasing a large subset of bibliographic data from Cambridge University Library catalogues as open structured metadata, testing a number of technologies and methodologies including XML, RDF, SPARQL and JSON.
Connecting Repositories - Open University
Making it easier to navigate between relevant scientific papers, using Linked Data format to describe the relationships between papers stored across a selection of UK Open Access repositories.
Contextual Wrappers - Cambridge University
Opening possibilities for discovering and using resources at the Fitzwilliam Museum, making these available through the Culture Grid, an aggregation service for museums, libraries and archives metadata.
Discovering Babel - Oxford University
Enhancing the resource discovery infrastructure of the digital literary and linguistic resources in the world-renowned Oxford Text Archive and in the British National



```
001 ocm32248821
003 OCoLC
005 19991030224027.0
008 950403s1994 xx b 001 0 eng d
040 $a TSW $c TSW $d OCL 043 $a n-us---
049 $a TSWB
100 1 $a Pullin, Michael Thomas, $d 1959- 245 12 $a A history of judicial
intervention in church property disputes / $c by Michael T homas Pullin.
260 $c c1994.
300 $a 2, ix, 220 leaves ; $c 29 cm.
502 $a Thesis (Ph. D.)--Southwestern Baptist Theological Seminary, 1994.
500 $a Includes abstract.
504 $a Includes bibliographical references (leaves 205-213) and index.
650 0 $a Church property $x United States $x Cases.
650 0 $a Church and state $x United States $x Cases.
650 0 $a Church controversies $x Cases.
```

Bibliographic data

Also could be Authority type information - Things like birth/date deaths



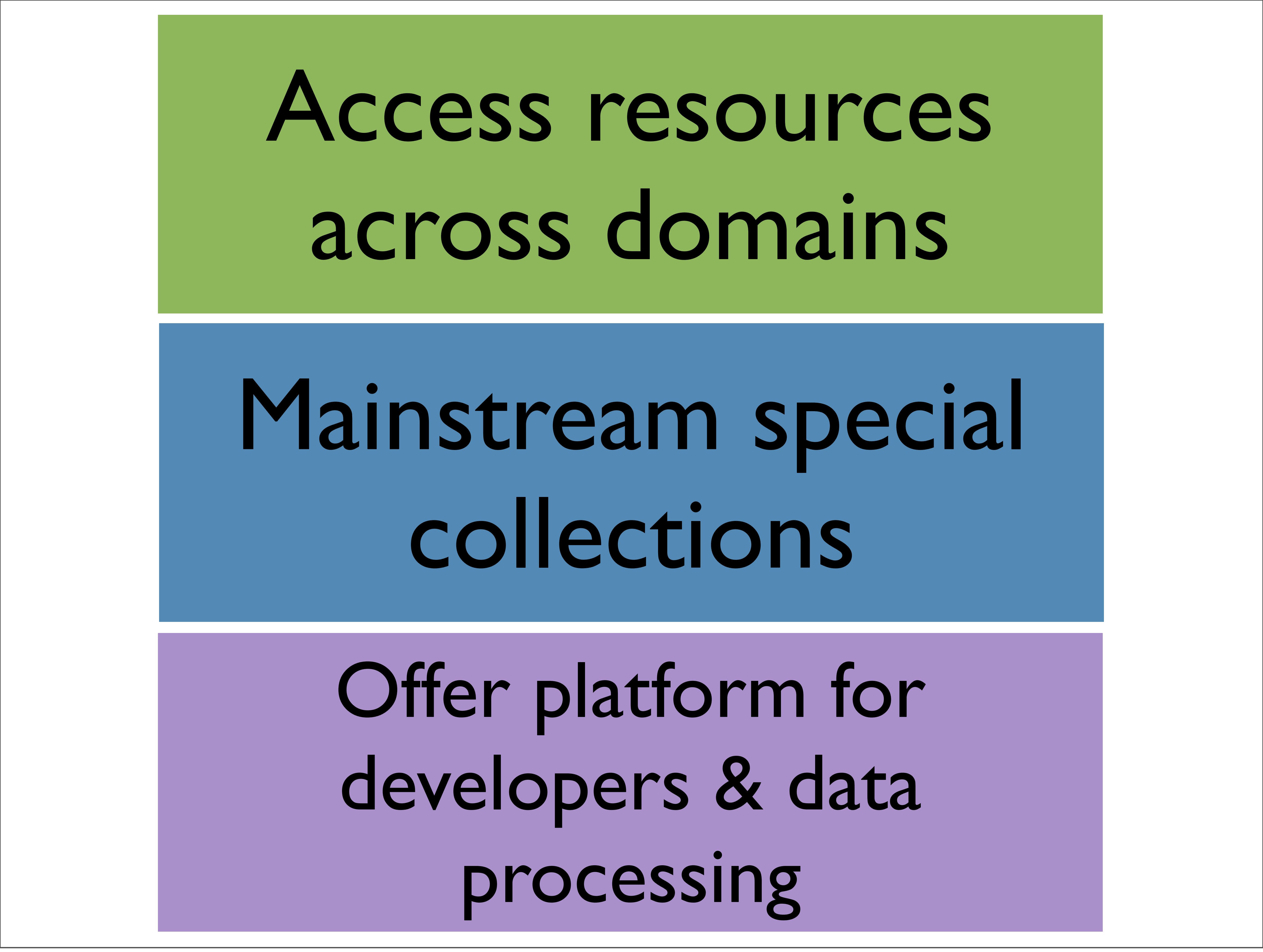
Ernest Henry Shackleton was born on 15 February 1874 in Kilkea, Ireland, one of six children of Anglo-Irish parents. The family moved from their farm to Dublin, where his father, Henry studied medicine. On qualifying in 1884, Henry took up a practice in south London, and between 1887 and 1890, Ernest was educated at Dulwich College. On leaving school, he entered the merchant service, serving in the square-rigged ship *Hoghton Tower* until 1894 when he transferred to tramp steamers. In 1896, he qualified as first mate, and two years later, was certified as master, joining the Union Castle line in 1899.

Excerpt from the Administrative History/Biography part of Finding Aid for the Sir Ernest Henry Shackleton Collection
http://archiveshub.ac.uk/search/record.html?id=gb15sirernesthenryshackleton

Aggregation

Wednesday, 30 March 2011

At centre of vision:
Currently working on aggregation ‘use cases’ - different concepts of aggregation with different issues – examples:



Access resources across domains

Mainstream special collections

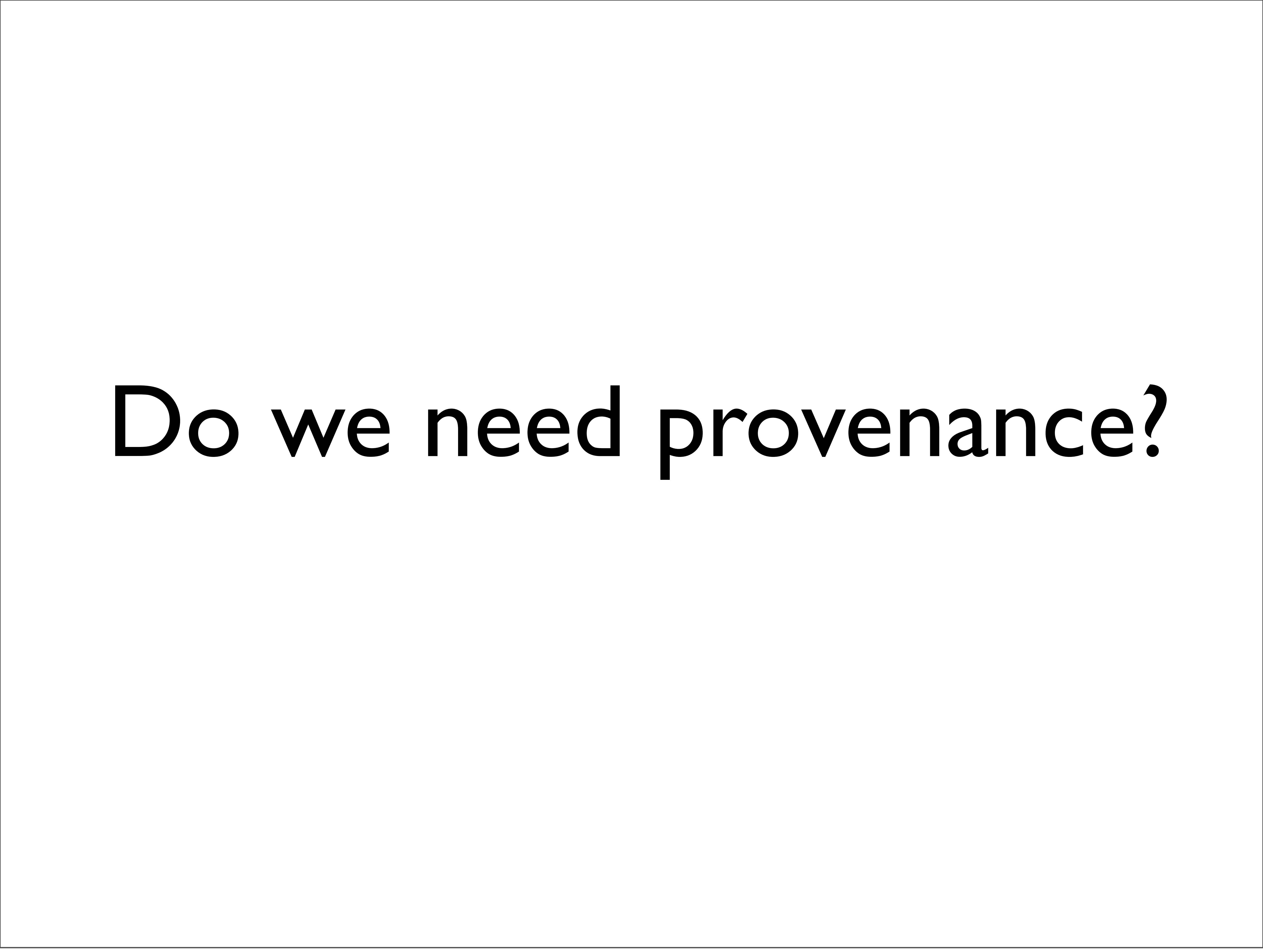
Offer platform for developers & data processing

Wednesday, 30 March 2011

3 Example Aggregation use cases: 8 projects strongly self-identified with this

Further Aggregation use cases...
User focussed:
•Mainstream Special Collections
•Use collection descriptions as an entry point
•Create a ‘first stop shop’
•Generate recommendation services
•Satisfy interdisciplinary enquiry
•Access resources across domains
•Serve the long tail
Aggregation focused:
•Create homogenous from heterogeneous
•Offer platform for developers and data processing
Data owner focused:
•Support collection management
•Data enhancement services
•Support catalogue record improvement
•Resource sharing across institutions
•Represent a Thematic Collection

Examples are … - and link to issues…

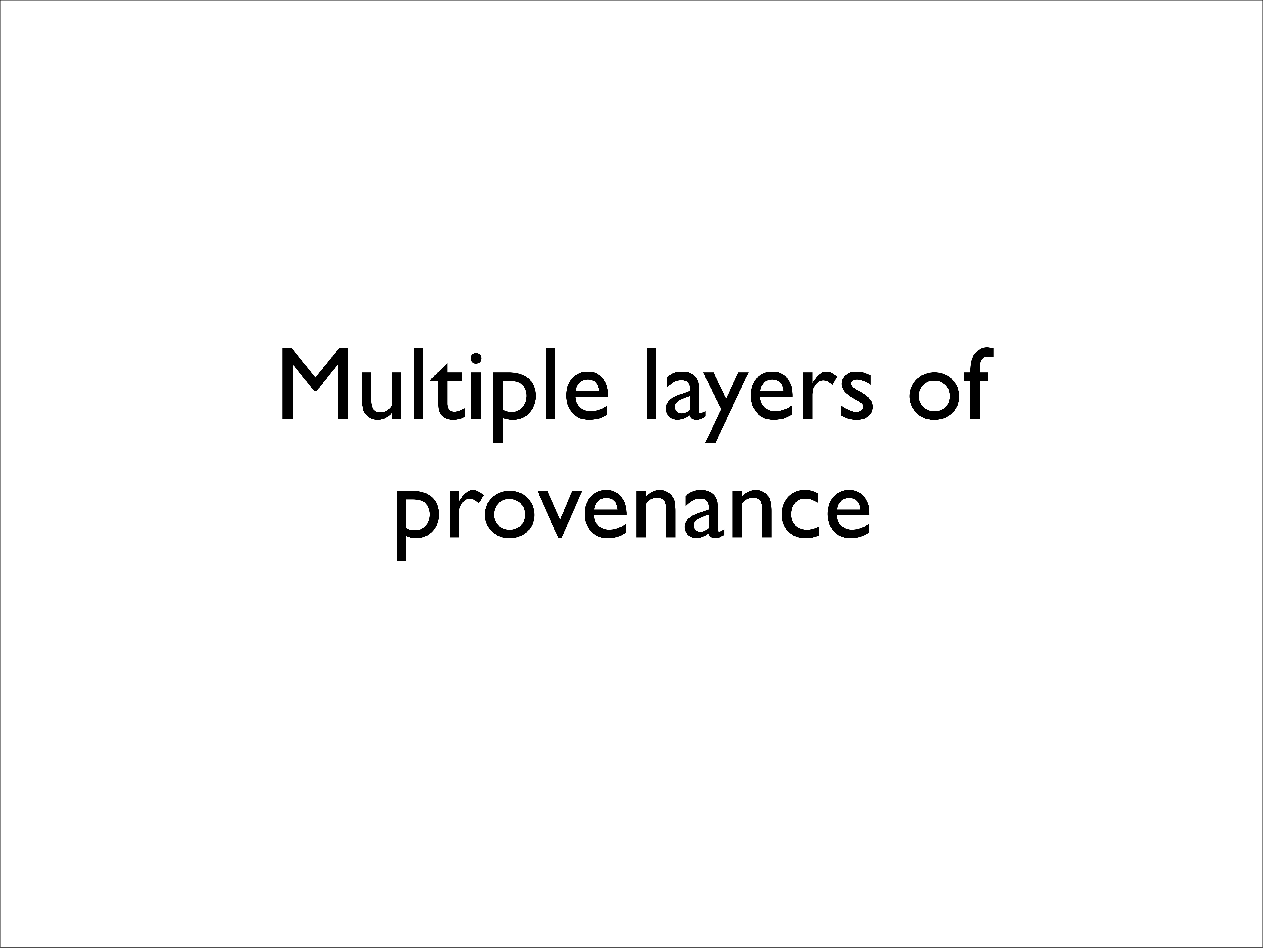


Do we need provenance?

Wednesday, 30 March 2011

Not simple answer - the provenance of a catalogue record for a contemporary mass-market paperback is probably not that interesting, but provenance for descriptions of rare of unique material is definitely of interest

Issues



Multiple layers of provenance

Provenance with respect to the sources, and provenance with respect to the aggregation

If you aggregate data – who vouches for provenance

Where do records come from? (and where do errors come from? Lucero experience!)

May include merging records; making assertions about the similarity of entities;

Derived data

Wednesday, 30 March 2011

Data derived from aggregation:
e.g. ‘related works’; ISBNx

Transformations

Multiple levels:
Source could transform before contributing to aggregation
Aggregation transforms

Changes to data models e.g. FRBR

Example of BNB -> Bibliographica (http://knowledgeforge.net/pdw/trac/wiki/datatriage)
Dates/Transcribed information such as place of publication – do aggregators and originator really understand the properties?

Example from Archives Hub – an existing aggregation, but not LD: ‘we don’t change the data because it is the contributors’ data’

Data owners often want reassurance that their contribution will be acknowledged – much much easier to agree cc-by that CC0

Terminology – ‘Provenance’ already taken in Museum space!

Terminology

Wednesday, 30 March 2011

Provenance already taken! Especially in Museums/Archives - Provenance used to talk about use of ‘provenance of objects in collection’ - could become confusing...

Licensing and Citation

Data owners often want reassurance that their contribution will be acknowledged - much much easier to agree CC-BY that CC0

Some source data may come from commercial providers
Some may come from academic research - e.g. RED/Lucero

Opening Doors, Opening Data
18th April, Manchester
http://rdtf.mimas.ac.uk

Wednesday, 30 March 2011

If you are interested in the work we are doing