The Past Is a Different Database – They Do Things Differently There.

Jeff Pache,
Product Development Manager,
IEE Inspec

IEE Inspec

www.iee.org/inspec

Background

- Most archiving is future proofing present data
- Most retro-digitisation is full text:
  - Page images + metadata
- This case study: Inspec's Backfile project:
  - Present-proofing the past
  - 100% metadata

Science Abstracts: 1898-1968

- 71 Years
- 176 Volumes
- 135,000 Pages
- 873,700 Abstracts
- 2,100,000 Index Entries
- Producing, via 25 GB of PDFs:

1 Database of 1.5GB XML + 3,675 Gifs

Differences of the Past

- Cultural
  - Different view of literature
  - Different view of people, e.g. “Mme P. Curie”
- Medium
  - Unstructured print vs structured XML
- Technological
  - Manual handling caused errors and gaps that computer validation would have prevented

Problems

- Obtaining raw material
  - A copy that can be destroyed in the process
- Knowledge of raw material
  - No amount of sampling will find all anomalies
- Curiosities
  - Handwritten tables
  - Multi-reference records

Problems Continued

- Anachronisms
  - Author treatment
- Bibliographic control
  - Lower and varying standards
- Printed page to structured record
  - “See previous abstract” references

Sources of Errors

1. Original printed data
   - Typographical errors and omitted data
2. Data capture process
   - Misreading, misinterpreting & miskeying
3. QC process
   - Automatic correction to deal with 1 and 2 can introduce further errors

Classification and Indexing

Problems:

- Variety of styles and formats:
  - 0 to 3 levels of subject headings
  - Class codes, no codes and UDC
- Archaic terminology
- Automatic application of modern terms might apply terms before their time

Classification and Indexing

Solutions:

- Map original indexing and classification to terms and codes from current Inspec thesaurus and classification
- Create additional “non-current” terms for extinct technology and theories

Conclusions

- Backfiles of significant age have a very different flavour from current raw data
- Understanding the data and the assumptions implicit in it is crucial
- Capturing the data in a computer file is but a small part of the process
- The more analysis, better the final product