Thirty problems for subject interoperability (and a few possible solutions)

Alan Dawson

Centre for Digital Library Research, University of Strathclyde, Glasgow G1 1XH

August 2006

alan.dawson@strath.ac.uk

This paper was presented at Electric Connections 2006

Abstract

Using evidence, and subject terms, drawn from the IRIScotland project, this presentation illustrates and summarises numerous problems that arise when aggregating metadata records from different sources. It aims to identify and classify the causes and consequences of these problems and suggest some ways of improving subject interoperability.

About IRIScotland

The Institutional Repository Infrastructure for Scotland project (IRIScotland) was 'conceived as a means to provide the organisational and technological framework for a Scotland-wide institutional repository infrastructure for research'. Project led by Edinburgh University:

http://www.iriscotland.lib.ed.ac.uk/

It will provide 'a harvester-based pilot cross-repository **search and browse service** to enhance exposure of the Scottish research output as a whole'. This component is led by CDLR at Strathclyde University.

So far, metadata from the repositories of five universities has been harvested and stored in a central database at Strathclyde University. Early pilot service available at:

http://cdlr.strath.ac.uk/iriscotland/

This presentation covers **subject terms** only, though there are interoperability problems with all metadata fields (titles, authors, dates, item types, journal names and issues, etc).

All examples are real, and drawn from a mere 5000 records.

Problem 1: Typos

Data entry errors - typing or spelling mistakes.

Example subjects from IRIScotland metadata:

hydro**go**logy
Motor **mo**ntrol
occ**ul**sion resolution
Schola**lr**y publishing-Africa
sp**ea**ch

Problem caused by human error and absence of controlled procedures.

Problem will affect searching and browsing.

Problem 2: Punctuation

Inclusion of punctuation symbols in subject field, e.g. commas, full stops, semicolons, question marks, exclamation marks.

Example subjects from IRIScotland metadata:

Antigenic variation, | genetic polymorphism,
homology manifolds.
Zoo conservation.
Theses Alive!
Toxocara; | Nematode; | Mucin |Glycoprotein
Eastern Mediterranean, | Neotethys, | Late Cretaceous,
Chromium Remediation or Release?

Problem caused by lack of guidance or training.

Problem may affect subject browsing.

Problem 3: Hyphens

Inconsistent hyphenation, inclusion of soft hyphens, M dashes, N dashes, double dashes.

Example subjects from IRIScotland metadata:

Bird-Meertens Formalism
Bird- Meertens Formalism
domain equa- tions
Multi-stage
Multicanonical
post-colonial

postinfection
tree - grass inter-cropping
3-D objects
3D vision

Problem caused by copying and pasting from documents, lack of standards and guidance.

Problem will affect searching and browsing.

Problem 4: Quotes

Apostrophes, single quotes, double quotes, smart quotes, back quotes, Macintosh quotes, Wingdings (often pasted in from documents).

Example subjects from IRIScotland metadata:

Alzheimer s

Alzheimerðżðð3/4ðð3/4(tm)s disease

Moggiðżðð3/4ðð3/4(tm)s computational metalanguage

Problem caused by Microsoft, Apple, by use of extended character sets, by copying and pasting, by lack of training and absence of checking procedures.

Problem will affect subject searching and browsing, but affects titles and abstracts more than subjects.

Problem 5: Spaces

Inclusion of double spaces or hard spaces or

Example subjects from IRIScotland metadata:

central nervous system
natural language
natural language
Tyne Estuary

Problem caused by human error and absence of checking procedures.

Problem will affect phrase searching and subject browsing.

Problem 6: ASCII

Inclusion of non-standard characters in subject terms.

Example subjects from IRIScotland metadata:

Búmodel
Hofstaðir
Sh??hn??ma illusrations

Problem caused by limitations of ASCII, use of extended character sets and lack of standards.

Problem will affect subject searching and browsing, but affects titles, author names and abstracts more than subjects.

Problem 7: Case

Mixture of upper case, lower case and mixed case terms.

Example subjects from IRIScotland metadata:

cognitive science
Cognitive science
Cognitive Science
Perception
perception
PUBLICATION BIAS, EVOLUTION, BEHAVIOR, ECOLOGY, KINSHIP, METAANALYSIS, SELECTION, TRIM, FILL
Ecology

Problem caused by human error, absence of checking procedures, standards and guidance.

Problem should not affect searching but may affect browsing.

Problem 8: Plurals

Mixture of singular and plural terms

Example subjects from IRIScotland metadata:

enzyme
enzymes
nematode
nematodes
vaccine
vaccines
typed lambda calculi
typed lambda calculus
NGO
NGOs
Institutional Repositories
Institutional Repository

Problem caused by absence of standards and guidance.

Problem should not affect searching but may affect browsing.

Problem 9: Places

Inclusion of place names as subject terms.

Example subjects from IRIScotland metadata:

Glen Affric
Oaxaca
Mexico
Sulawesi
Quebec
Baffin Island
Tweed Estuary
Gabon
Zimbabwe

Problem caused by absence of standards and guidance.

Problem will affect subject browsing.

Problem 10: People

Inclusion of personal names as subject terms.

Example subjects from IRIScotland metadata:

Barthes
Derrida
Gavin Hamilton
Handelman
Lerdahl
Lord Jesus Christ
Christine de Pizan
Park, Edwards Amasa, 1808-1900
Edwards Amasa Park
Jonathan Edwards
Jackson, Frank, 1943-

Problem caused by absence of standards and guidance.

Problem will affect subject browsing.

Problem 11: Organisations

Inclusion of corporate bodies as subject terms.

Example subjects from IRIScotland metadata:

Andover seminary
University of Asmara
University of Edinburgh
Edinburgh University
EUL
Global Analysis, Integration and Modeling Task Force
United Kingdom s Medical Research Council
Lothian health board

Problem caused by absence of standards and guidance.

Problem will affect subject browsing.

Problem 12: Coverage

Inclusion of geographical coverage qualifiers within subject terms.

Example subjects from IRIScotland metadata:

Aquaculture Chile
British cattle
Pop music Portugal
Corporations, Taiwanese Great Britain Personnel management
Aquaculture Sri Lanka
House selling Scotland North Lanarkshire

Problem caused by absence of standards and guidance.

Problem will affect subject browsing.

Problem 13: Projects

Inclusion of project names as subject terms.

Example subjects from IRIScotland metadata:

OCEAN DRILLING PROGRAM LEG 180
ONOMASTICA project
Sloan Digital Sky Survey
Hill Sheep and Native Woodland project

Problem caused by absence of training and guidance.

Problem will affect subject browsing.

Problem 14: Types

Inclusion of item type in subject field.

Example subjects from IRIScotland metadata:

Phd thesis
PhD thesis
PhD | thesis

Problem caused by absence of training and guidance.

Problem will affect subject browsing.

Problem 15: Abbreviations

Use of acronyms and abbreviations along with, or instead of, the full phrase.

Example subjects from IRIScotland metadata:

AI
Artificial intelligence
ASR
Automatic speech recognition
geographic information systems
gis
Proliferative kidney disease (PKD)
PKD (Proliferative kidney disease) proliferative kidney disease
embryonic stem (ES) cells
virtual learning environments
VLE

Problem caused by absence of standards and guidance.

Problem may affect phrase searching and will affect browsing.

Problem 16: Synonyms

Use of synonyms or variations of the same word or phrase.

Example subjects from IRIScotland metadata:

climate change
climatic change
dementia
demented
mud volcanism
mud volcanoes
Neural networks (Computing science)
Neural Nets

Problem caused by absence of standards and guidance.

Problem will affect searching and browsing.

Problem 17: Spelling

Mixture of English and American spellings.

Example subjects from IRIScotland metadata:

body odors
body odours
palaeography
paleography
Tumors
antitumour complexes
computational modeling
Ecological modelling

Problem caused by absence of standards and guidance.

Problem will affect searching and browsing.

Problem 18: Articles

Inclusion of definite or indefinite article in subject field.

Example subjects from IRIScotland metadata:

the event
the right to buy
the temporal coherence

Problem caused by absence of standards and guidance.

Problem may affect subject browsing.

Problem 19: Phrases

Inclusion of long phrases, short phrases and individual words as subject terms.

Example subjects from IRIScotland metadata:

capillary electrochromatography
Capillary
Electrochromatography
time-of-flight mass spectrometry | two-step laser mass spectrometry
Time-of-flight | Mass | Spectrometry
lumbriculus variegatus, feeding, toxic anorexia, in situ bioassay

Problem caused by absence of standards and guidance.

Problem will affect subject browsing.

Problem 20: Latin

Mixture of common names and Latin names.

Example subjects from IRIScotland metadata:

Babyrousa babyrussa / babirusa
African catfish, Clarias gariepinus (Clarias gariepinus)
Antirrhinum majus | snapdragon
Pinus sylvestris | native pinewoods

Problem caused by absence of standards and guidance.

Problem may affect searching and will affect browsing.

Problem 21: References

Inclusion of cross-references in subject field.

Example subjects from IRIScotland metadata:

N Visual arts (General) For photography, see TR

Problem caused by absence of guidance.

Problem will affect searching and browsing.

Problem 22: Keywords

Mixture of uncontrolled keywords along with controlled subject terms.

Example subjects from IRIScotland metadata:

anatomy
QM Human anatomy
Cancer
RC0254 Neoplasms
scale | granularity | hierarchy | process | mereology | ontology | geography
smolder | smoulder | polyurethane | space | fire | kinetics | modelling | modeling
Aged Institutional care Great Britain | Old age homes Great Britain | Quality of life Great Britain | Good Practice, Senior Citizens, Quality of Life, Ageing, Residential Care, Models of Ageing, Medical Model

Problem caused by absence of standards and guidance.

Problem will affect searching and browsing.

Problem 23: Context

Use of adjectives or abstract nouns that are almost meaningless out of context.

Example subjects from IRIScotland metadata:

changes
command
control
preference
principal
reduced
response
restriction
states
system
young

Problem caused by absence of standards and guidance.

Problem will affect browsing and retrieval precision.

Problem 24: Ambiguity

Use of subject terms that are meaningful but ambiguous out of context.

Example subjects from IRIScotland metadata:

Aids for the Deaf
Bands
Mansfield
stocking enhancement

Problem is inherent in use of language and caused by absence of taxonomy and authority control.

Problem will affect searching and browsing.

Problem 25: Identifiers

Inclusion of subject identifier in subject field.

Example subjects from IRIScotland metadata:

QA Mathematics
QA75 Electronic computers. Computer science | TP Chemical technology
HT Communities. Classes. Races | RA0421 Public health. Hygiene. Preventive Medicine | QP Physiology
ZA4050 Electronic information resources

Problem caused by software used and procedures followed in repositories.

Problem will affect subject browsing.

Problem 26: Classification

Use of a broad subject classification scheme to provide subject terms will be of limited value for item retrieval if the same term is applied to large numbers of records.

Example subjects from IRIScotland metadata:

QC Physics (903 records - 18%)
RK Dentistry (247 records - 5%)
QA76 Computer software (183 records - 3.7%)

Problem caused by lack of precision in chosen standard (LCC - Library of Congress Classification). LCC does not attempt to evaluate the subject content of items, but broadly categorises items in a subject hierarchy. The degree of precision required depends on the nature and scope of the service being provided.

Problem will affect searching and browsing.

Problem 27: Topics

Use of very specific topics as subject terms merely repeats words that occur in a title or abstract. This makes the subject term redundant and renders meaningful subject browsing almost impossible.

Example subjects from IRIScotland metadata:

3-dehydroquinase
AMPA/kainate glutamate receptors
monofunctional organometallic ruthenium (II)
aryl hydrocarbon receptor nuclear translocator
Mixtures of multivariate Bernoulli distributions
Ube4b/Nmnat
hypothalamic oxytocin neurons
Girard-Tait reducibility

Problem caused by absence of standards and guidance.

Problem will seriously affect subject browsing.

Problem 28: Structure

The software in common use for repositories (Eprints and DSpace) encourages several subject terms to be included in a single field, separated by delimiters. This is hopeless for subject browsing.

Example subjects from IRIScotland metadata:

counselling | feminisms | policy | UK | women
Kinetics | micelle | Aqueous lecithin-bile salt mixtures | vesicle
Spoken document retrieval | Information retrieval | Broadcast speech
Spoken document retrieval | Preservation | Copyright | Speech technology

stability of tall buildings | collapse mechanism | WTC Towers | Structure in fire
Stable oxygen isotope | Eratosthenes Seamount | surface-water conditions

Problem caused by software and procedures.

Problem will seriously affect subject browsing.

Problem 29: Absence

Absence of any subject terms at all

Example subjects from IRIScotland metadata:

Problem affects 456 out of 5000 records (9%)

Problem caused by procedures and absence of guidance.

Problem will affect searching and browsing.

Problem 30: Schemes

Use of different subject schemes (taxonomies, thesauri, vocabularies, ontologies).

Example subjects from IRIScotland metadata:

Birds Great Britain Geograpical distribution | Birds Scotland Geographical distribution
Birds|zScotland|zCentral Scotland | Woodland|zScotland|xBirds

Problem potentially caused by use of different subject schemes in different institutions. In practice the only controlled scheme in use is LCC (by Glasgow, St Andrews and Strathclyde universities) so the problem of scheme interoperability has not yet arisen. Edinburgh and Stirling universities use uncontrolled keywords.

The issue of mapping between different subject schemes is being investigated by the HILT project at CDLR.

Problem will affect searching and browsing.

Problem 31: Languages

Use of subject terms in different languages.

Not applicable to IRIScotland (yet)

Problem will affect searching and browsing.

Problem summary

All these problems have been observed in a mere 5000 *mediated* records from just five universities, all describing one type of resource: academic articles, papers and theses.

Problems with a relatively easy solution - clear procedures and guidance

Typos
Punctuation
Hyphens
Quotes
Spaces
Case
Places
People
Organisations
Projects
Types
Articles
References
Identifiers
Structure
Absence

Problems with a more difficult solution - requiring use of agreed standards

Plurals
Abbreviations
Synonyms
Spelling
Latin
ASCII
Keywords

Problems with fairly difficult solutions - dependent on scope of service

Coverage
Ambiguity
Context
Classification
Phrases
Topics
Schemes
Language

Some solutions

Guidance for staff applying subject terms

- Don't include punctuation
- Don't include 'the'

- Don't include cross-references
- Don't copy and paste quotes, hyphens or dashes
- Don't use names of people, places, corporate bodies or projects as subject terms
- Don't use meaningless terms
- Do use at least one subject term
- Do use lower case for subject terms
- Do put subject terms in separate fields

Some solutions

Procedures for service providers

- Define and apply metadata cleaning procedures
- Remove double spaces
- Remove punctuation
- Apply character translations
- Convert subject terms to lower case
- Remove or separate subject identifiers
- Produce reports of subject terms used, to easily identify anomalies
- Define and agree metadata profile covering semantics (content) as well as syntax (structure)

Guidance for resource creators

- Include the names of any significant people, places, species, projects or topics in the item abstract (especially if not in the title)
- Avoid keyword clutter

Some big problems for IRIScotland

- Use of uncontrolled keywords and topics and subject terms
- Lack of a suitable single broad subject scheme, like LCSH but British and better
- Lack of standards for qualified Dublin Core, e.g. to distinguish subjects from people and places
- Lack of authority control for people and place names
- Full text searching not available
- Lack of control over procedures at contributing institutions

Conclusions

- Subject interoperability is easy in principle but difficult in practice
- Many problems can be solved relatively easily, by improved procedures and clear guidance
- Service providers should define and apply metadata cleaning procedures
- Service providers should try to define and agree metadata profiles, to encourage standardisation and interoperability
- Uncontrolled keywords may have some value for searching, but are disastrous for subject browsing

- Vagaries of language and meaning pose intractable problems, but most interoperability problems have much simpler causes
- Some solutions depend on the size and scope of the service being offered and the level of precision required for item retrieval