

Language independent metadata

Juha Hakala, Helsinki University
Library, Finland

ELAG 2003

Bem, 2-4 April 2003

Content

- Definitions
- Multilingual metadata in MARC21 context
 - other metadata formats will not be discussed!
- Global authority control
- Virtual International Authority File
- MARC and multilinguality
- Conclusions and recommendations

Definitions

- Cross-Language Information Retrieval (CLIR) is generally divided by researchers in:
 - bilingual IR (all documents in one language, search in any other language)
 - multilingual IR (documents and metadata in multiple languages)
- Three methods:
 - thesaurus/dictionary based; corpus based; machine translation (authority records not included!)
- Inherently multilingual metadata (codes, classifications) is outside the scope of IR research

Definitions (2)

- Language independent metadata (or a system) supports bi- or multilingual IR
 - bilingual access : metadata itself is language independent (e.g. MARC21 codes, classifications)
 - multilingual access: metadata is catalogued in multiple languages or is converted on the fly to multiple languages using thesauri, corpus, authority files or machine translation

Bi- and Multilingual metadata in MARC21 context

- MARC data is still primarily aimed at human users (production of catalogue cards), not for advanced automated processing
 - ISBD punctuation in data!
- MARC is machine readable, not understandable
- Attempts to make MARC 21 more digestible for machines (such as increased reliance on codes) tend to make it harder for cataloguers to use, and have therefore been criticised

Multilingual metadata in MARC21 context (2)

- Shift from the use of classifications to utilisation of subject headings
 - a problem for e.g. foreign users of the Finnish national bibliography
 - There is an urgent need for cross lingual mappings between different subject headings
- Classifications used in the past and subject headings utilised at present should be linked

Multilingual metadata in MARC21 context (3)

- Development of multilingual thesauri
 - requires sophisticated ontology development; emerging interest in Semantic Web community
 - Finnish Subject Headings List -> Finnish General Ontology
- Provide authority control on the global scale
 - large number of recent initiatives, such as LEAF, Interparty
 - efficient but time consuming approach; slim academic (research) interest

Global authority control

- First step: match existing authority files
 - aim: match existing records of the same entity
 - automated matching developed e.g. in OCLC will help, but a lot of manual effort will be needed as well
- Programmatical creation of links to preferred entity names

Global authority control (2)

- The result of automated and manual matching will be an exhaustive authority record for an entity
- An example provided by Barbara Tillet (from A Virtual International Authority File –presentation, 22.11.2002)
 - same entity with multiple variant scripts
 - unfortunately the example will lack the Finnish name form of the entity in question...

Tag	I1	I2	Subfield Data
010			#a n 80050515
035			#a (DLC)n 80050515
040			#a DLC #c DLC #d DLC #d NIC
100	0		#a Confucius
400	0		#a Konfuzius
400	0		#a K'ung Fu-tzu
400	0		#a Kongzi
400	1		#a Kong, Qiu
400	0		#a K'ung-tzu
400	1		#a K'ung, Ch'iu
400	0		#a K'oshi
400	0		#a Konfu ^ˆ t ^ˆ si ^ˆ i
400	0		#a Kongja
400	0		#a Kung Fu
400	1		#a K'ung, Fu-tzu
400	0		#a Confucio
400	0		#a Конфуций
400	0		#a 孔夫子
400	0		#a 孔子
400	0		#a 孔丘
400	0		#a こうし
400	0		#a コウシ
400	0		#a 공자
670			#a Jakobs, P. M. Kritik an Lin Piao und Konfuzius, c1983: #b t.p. (Konfuzius)
670			#a Konfu ^ˆ t ^ˆ si ^ˆ i, 1993: #b t.p. verso (551-479 B.C.)
670			#a His Gespr ^ˆ ache (Lun y' u), 1910: #b t.p. (Kungfutse)
670			#a Web connection #u http://www.friesian.com/confuci.htm
700	0		#a 孔夫子 #5 Natl. Lib. of China

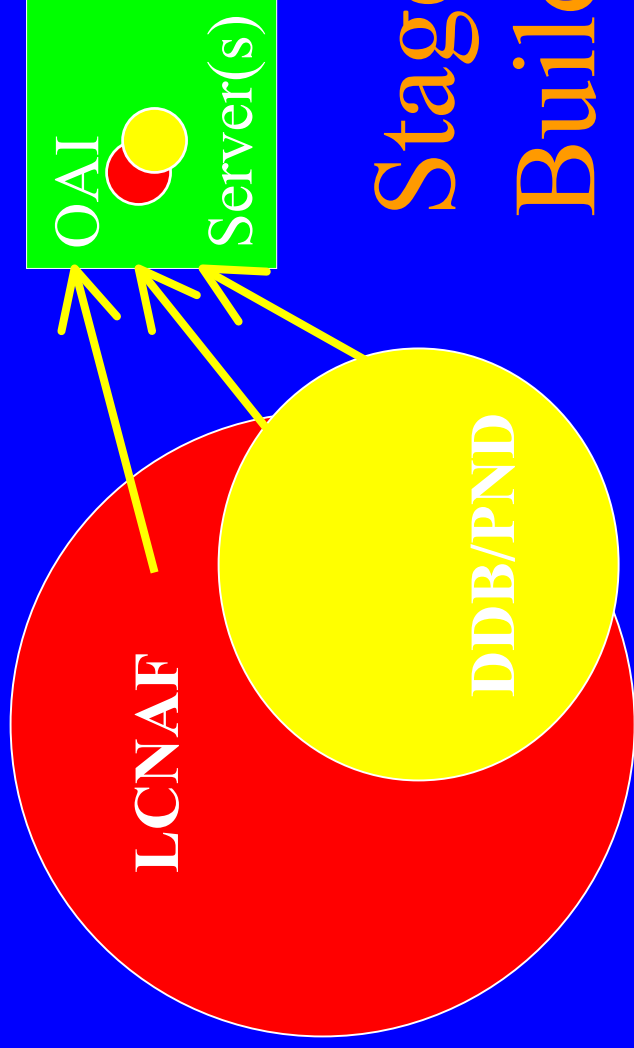
Virtual International Authority File: implementation options

- Independent national/regional authority files searched via Z39.50 (Bath profile v. 2)
 - this model is not scaleable beyond 10-15 authority files
- Simultaneous Z39.50 search using International Standard Authority Data Number
 - scales rather well, but requires lots of processing of incoming data plus global implementation of ISADN, which is non-trivial requirement
- Global authority file created with OAI harvesting

VIAF proof of concept

- Library of Congress, Die Deutsche Bibliothek and OCLC will test the centralised union authority file model
- OAI will be used for harvesting the metadata
- Description of the project (again by Barbara Tillett)

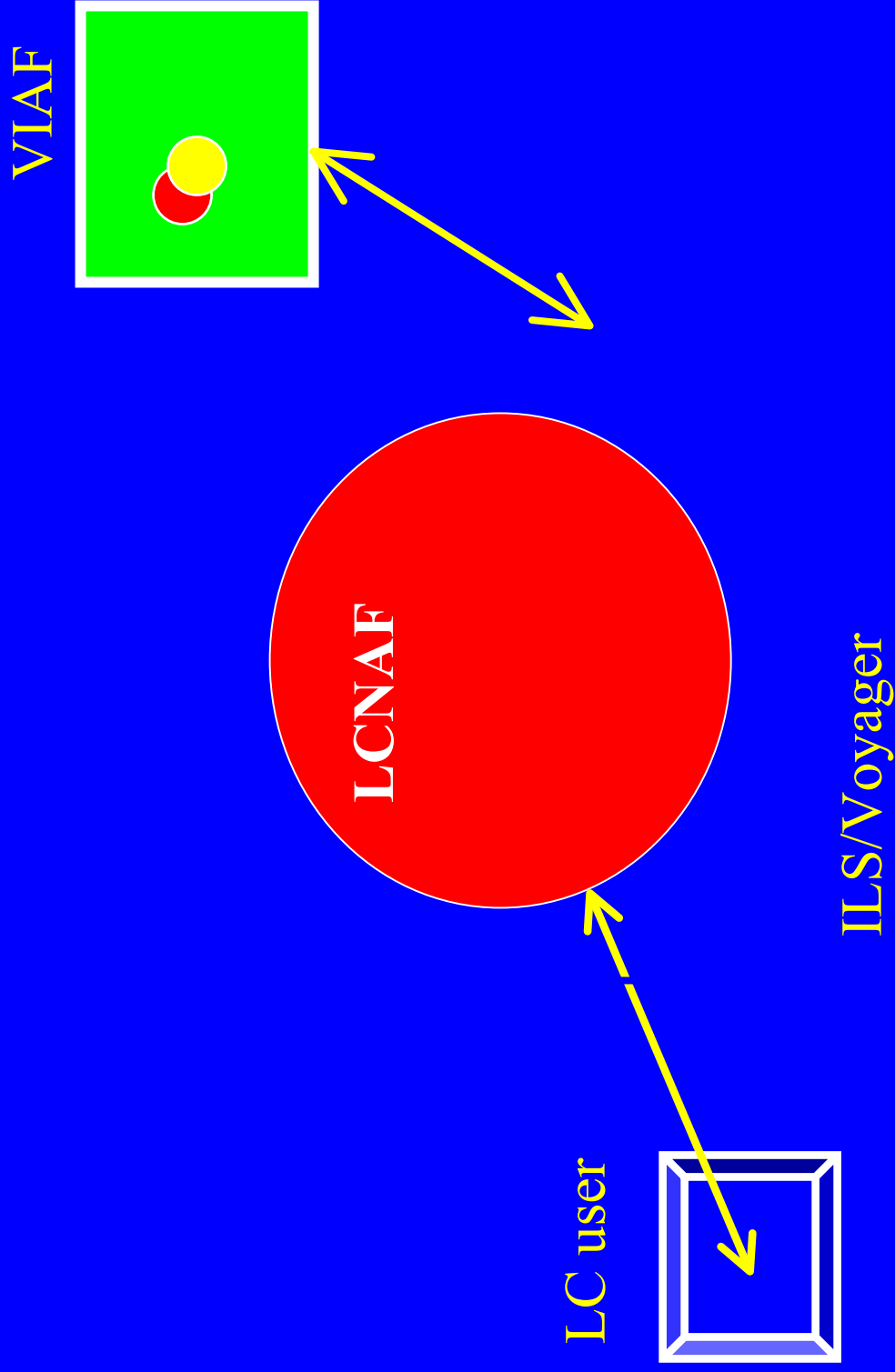
VIAF Proof of Concept DDB/LC/OCLC



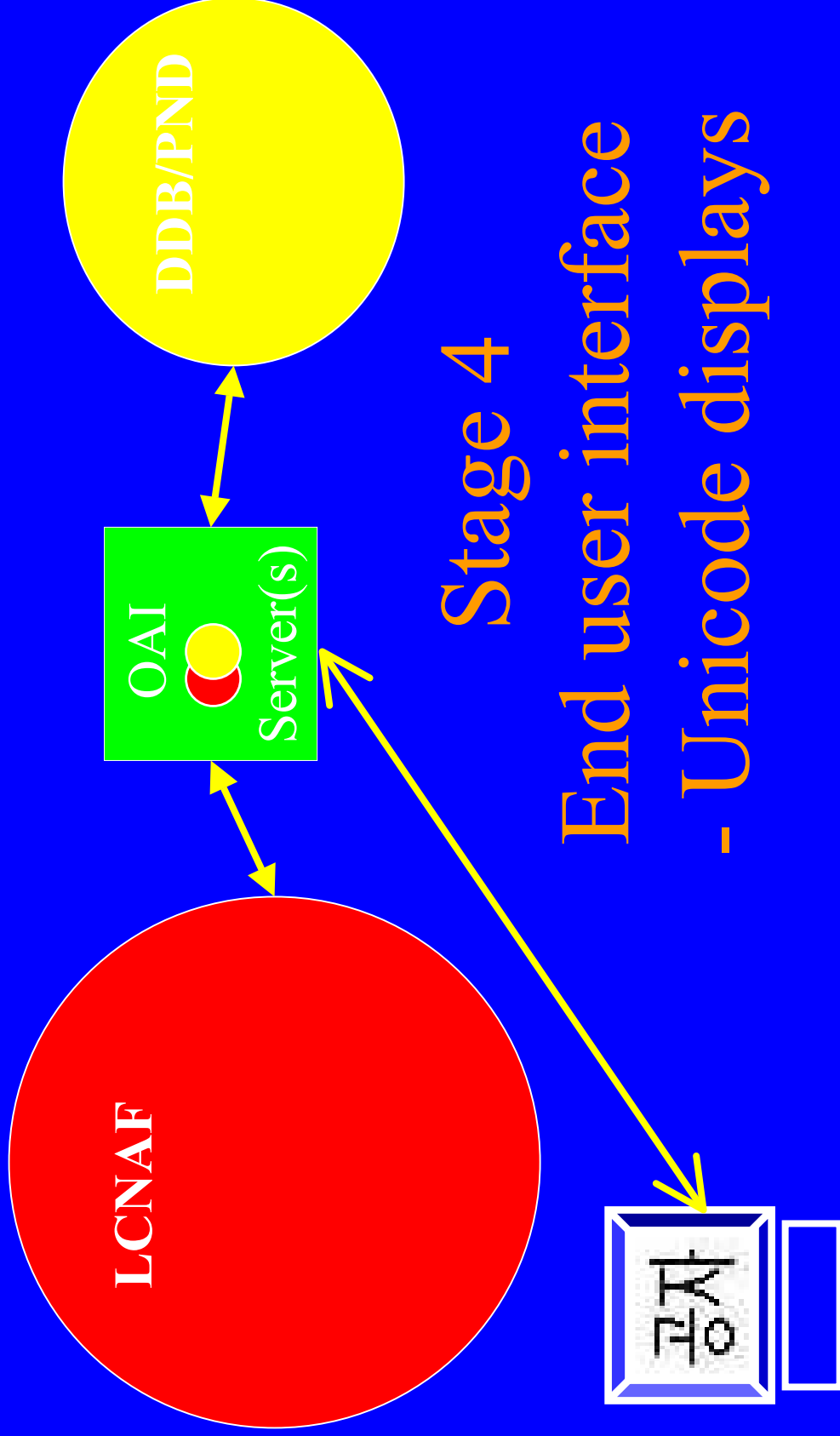
Stage 2
Build OAI
Server

VIAF Proof of Concept

DDB/LC/OCLC



VIAF Proof of Concept DDB/LC/OCLC



MARC21 and multilinguality: authority data

- MARC21 Format for Authority Data (and cataloguing rules) provides a good starting point for building multilingual applications
 - covers names and subjects
- In practice, due to lack of Unicode support and international cooperation (or tools facilitating it) our present (name) authority files and means for sharing data are very far from perfect
- VIAF and related projects will improve this!

MARC21 and multilinguality: bibliographic data

- Many tags have built-in support for multilingualism (e.g. 245, 250)
 - but the language may not be indicated and it is not possible to select only one display language
 - and cataloguing rules set strict limits for translation
- Some tags are dependent on cataloguing language (300); no room for parallel languages
 - machine translation by ILS possible for display
- Code fields such as 006-008 are "immune"; they must be processed for (meaningful) display

MARC21 and multilinguality: bibliographic data (2)

- Increased reliance on copy cataloguing has led to the relaxation on rules concerning the cataloguing language
 - 95 % of new acquisitions copy catalogued in HUL; language is not altered (except by some purists; there is no requirement for this)
- However, data should be multilingual (encoded) if it really counts

MARC21 format: holdings data

- Some tags are code driven – simple translation in application level to basically any language; e.g. in 863-865:
 - 21 = Spring / Frühling / Vår / Kevät
 - 01 = January / Januari / Tammikuu
- 866/867/868 and other non-structured free text tags are a problem, like note subfields in otherwise encoded fields

MARC21 format: general conclusions

- Difficulty of supporting multilingual access with present rules and MARC formats varies from impossible to difficult to piece of cake, depending on the case
- Given the background of AACR2 and MARC (production of printed bibliographies in the U.S.A.) this is not too surprising

Multilinguality, ILS and portals

- We need systems which allow multilingual access based on patron identification
 - based on identification, a portal must be able to retain the proper GUI language and process the authority, bibliographic and holdings records for display purposes accordingly (possibly using machine translation)
 - more transient methods such as cookies (stored after language selection) can also be used if patron level authentication is out of question

Multilinguality, MARC21 and cataloguing rules

- We need rules and formats which do not prevent doing things which make sense from bi- and multilingual access point of view
 - more widespread and systematic use of codes
 - possibility to provide translations (from beyond the publication itself) when needed
- More investment of authority and ontology work!
 - Multilingual name authorities for uniform and series titles, authors; cross lingual subject headings