National Library of Finland

Digitisation Policy, Production Processes and Access

(...and beyond Access)

Tiina Ison, Senior Analyst

Present-Day Library: Space, Design, Resources
Helsinki 25.3.2011
Russian Library Heads – NLF visit

Giuseppe Acerbi

Travels through Sweden, Finland and Lapland to the North Cape
in the years
1798 and 1799

Euroopan unioni
Euroopan aluekehitysrahasto
Euroopan sosiaalirahasto

Table of Contents

1. Digitisation Policy
2. Contextualization – the workflows….
3. Digitisation Production - containers ….
4. Digitisation Production - content….
5. Access to Digital Collections…
6. Beyond Access…
7. Experimentation with Beyond Access…

1. Digitisation Policy National Library, Finland

http://www.kansalliskirjasto.fi/attachments/5v5daJ8e3/5uhdIBk6X/Files/CurrentFile/NLF_Digitisation_Policy.pdf

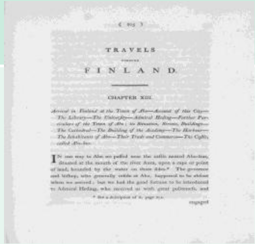
New information resources created through library digitisation:

1. **Creation of digital manifestation of source material** *(i.e.book)*
2. **Enabling content granulation via structural mark up** *(i.e. chapter)*
3. **Creation of digitized corpus of text for automated text extraction** *(i.e. concepts/named entities in text->ontologies)*
4. **Creation of new metadata and enrichement of metadata through digitisation processes**
5. **Enabling crowd sourcing and tagging of content in digitisation production** *(i.e. distributed workflows for content /sttructural mark-up/annotation of themas/semi-automated processes)*
6. **Ensuring sustainable digitisation via life cycle management and digital preservation of archival copies** *(i.e. specs for each material type/treatement of aggregates(ER verus OO)/ how to update archival files)*

http://www.nationallibrary.fi/libraries/dimiko/digitisationpolicy.html

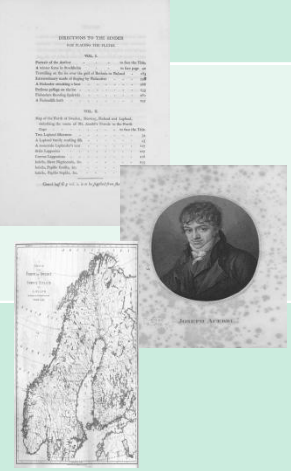
2. Contextualization – the workflows….

Work	Work	Work	Granularity (Fragmentation)	Authenticity Provenance LOD - URI
Acerbi; Physical Manifestation 1	Acerbi Digital Manifestation 2	Acerbi Digital Aggregates of Manifestation 2		Persistent Identifiers and links
			Structural Fragmentation Chapter Paragraph, Word, Character	Persistent Identifiers and links
			Conceptual Models Ontologies,Themas Relations Named Entities	Persistent Identifiers and links


TRAVELS

FINLAND

CHAPTER XIII



(Catalogue)
Container
Long
Library
Workflow
Tradition

(METS) Containers

(Semantic) Content

How to integrate new (and distributed) workflows in digitisation for contextualising and extracting meaning ?

3. Digitisation Production - MARC Container

Library Catalogue !

Minimal Record for
Digitisation Workflow
Unique ID of Source
Provenance of Source
(un)Controlled (linked) Vocabs

Value Drivers;
Authenticity of Source
and Links

Creator: Acerbi, Giuseppe, 1773-1846.

Title: Travels through Sweden, Finland, and Lapland, to the North Cape, in the years 1798 and 1799 / by Joseph Acerbi. In two volumes. Illustrated with seventeen elegant engravings. Vol. I[-II].

Publisher: London : Printed for Joseph Mawman ... , 1802 (by T. Gillet, Salisbury Square)

Format: 2 vol. (xxiv, 396 p., [8] leaves of plates ; viii, 380 p., [9] leaves of plates, some col.) : ill., map ; 4:o.

Note Vol. I: Portrait of Joseph Acerbi. Painted by P. Violet, engr. by P. W. Tomkins.

Vol. II: Music p. 325-336.

Handwritten notes by James Henry Monk (1784-1856) in volume I, identified by Maurice de Coppet. See also p. 92.

Provenance: Bequeathed by Maurice de Coppet (1868-1930), French ambassador to Finland in 1923-1929. Bookplate "Gallia in Finlandia", designed by the Finnish artist Jukka Pellinen, added later by the National Library of Finland.

Aineisto: kirja

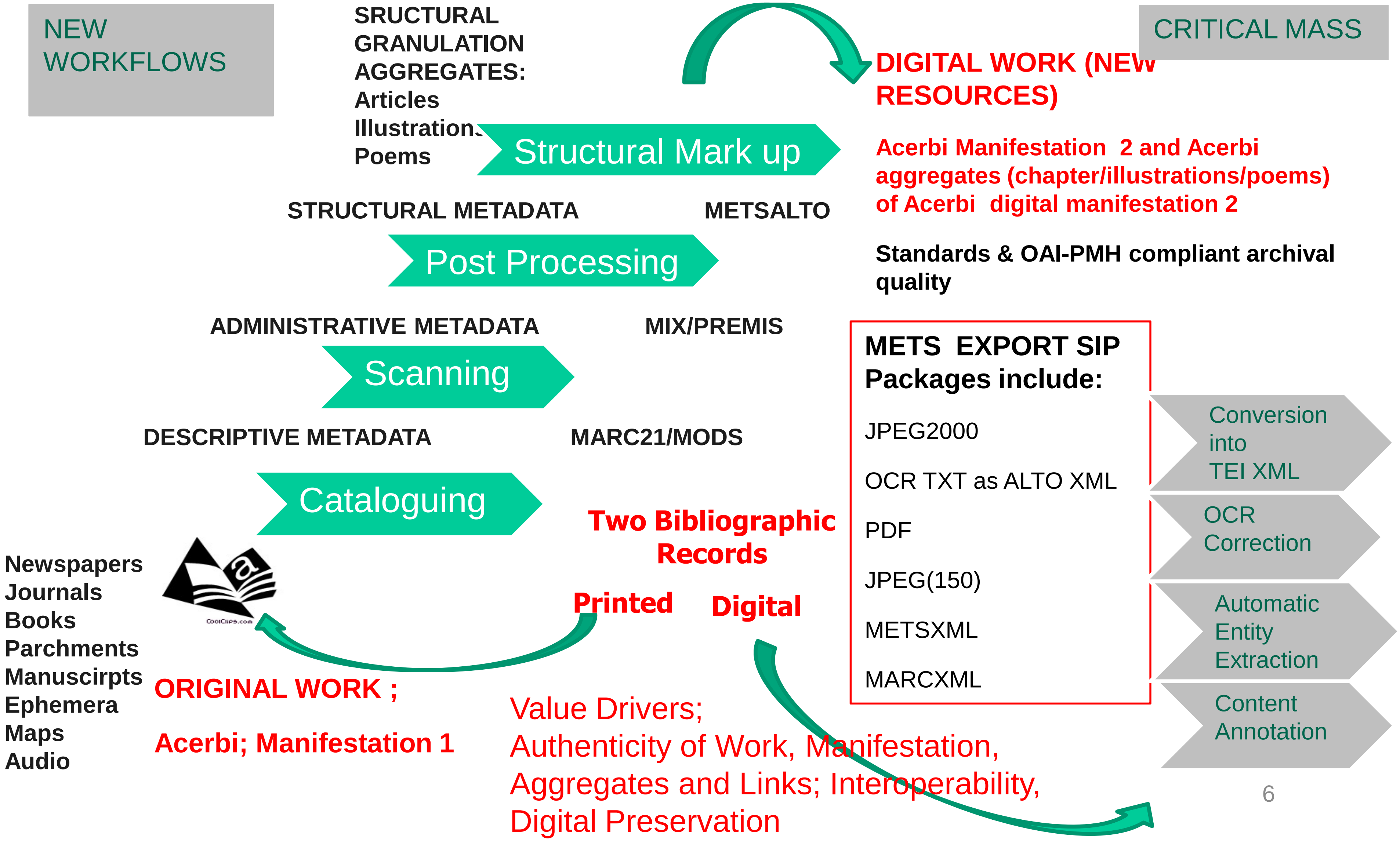
Language: eng

Subject: travel accounts
Sweden
Finland
Lapland

Abstract: Giuseppe Acerbi, Italian naturalist, explorer and composer, travelled to the far North of Scandinavia in the years 1798 and 1799. His travel account was first published in English and it was soon translated into several languages.

YSO – Upper Finnish Ontology
(English). Eero Hyönen Semantic
Web. FinOnto tool

3. Digitisation Production - METS Containers



NEW WORKFLOWS

SRUCTURAL GRANULATION AGGREGATES:
Articles
Illustrations
Poems

Structural Mark up

STRUCTURAL METADATA METSALTO

Post Processing

ADMINISTRATIVE METADATA MIX/PREMIS

Scanning

DESCRIPTIVE METADATA MARC21/MODS

Cataloguing

Newspapers
Journals
Books
Parchments
Manuscirpts
Ephemera
Maps
Audio

ORIGINAL WORK ;

Acerbi; Manifestation 1

Two Bibliographic Records

Printed Digital

Value Drivers;
Authenticity of Work, Manifestation,
Aggregates and Links; Interoperability,
Digital Preservation

CRITICAL MASS

DIGITAL WORK (NEW RESOURCES)

Acerbi Manifestation 2 and Acerbi aggregates (chapter/illustrations/poems) of Acerbi digital manifestation 2

Standards & OAI-PMH compliant archival quality

METS EXPORT SIP Packages include:

JPEG2000

OCR TXT as ALTO XML

PDF

JPEG(150)

METSXML

MARCXML

Conversion into TEI XML

OCR Correction

Automatic Entity Extraction

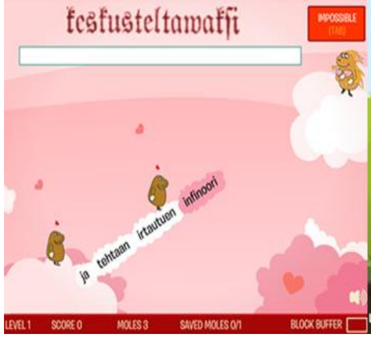
Content Annotation

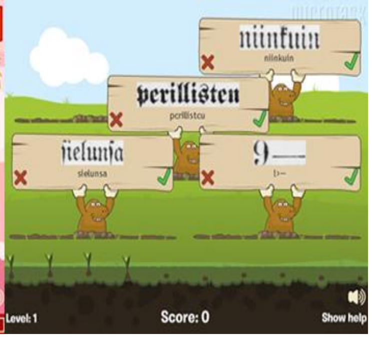
4. Digitisation Production – OCR correction….

National Library of Finland and Microtask, 2011

http://www.microtask.com

OCR Correction of Historical Digitized Newspapers via crowd sourcing and gaming ‘Mole Hunt’, players are shown two different words from which they must determine any similarities. ‘Mole Bridge’ makes players correctly spell the words appearing on the screen.





- ALTO files generated via digitisation OCR processes
- OCR recognition of old characters, Fractur poor.
- Poor quality obstacle for automated extraction of i.e. named entities.
- OCR correction is performed via Microtask gaming, experimental project.
- Text can be converted to TEI/XML

Next mark up of articles/illustrations in historical digitised newspapers and digitsed journals through gaming ….

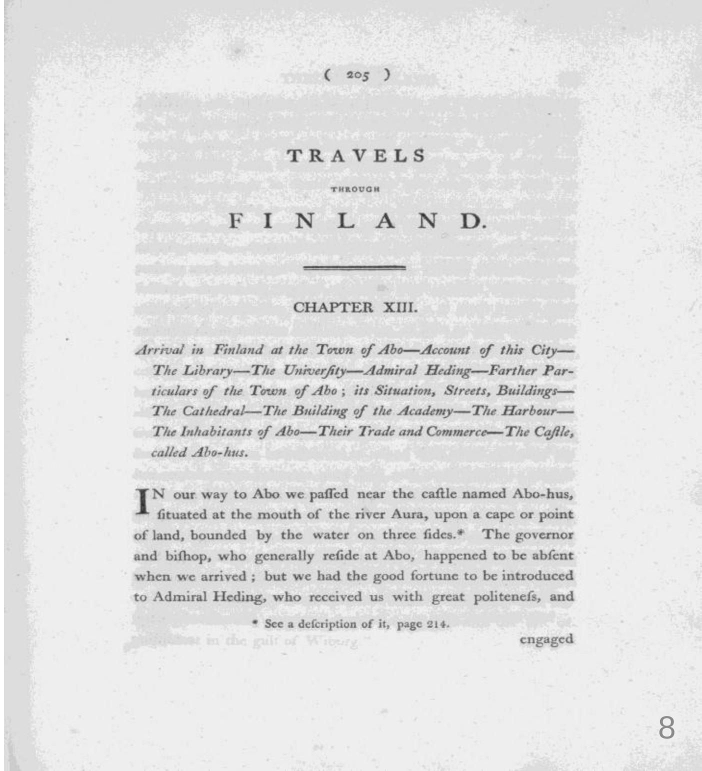
More than 25 000 people have completed over 2 million individual tasks inside the games. This measures up to about 100 000 minutes or about 1700 hours of work (or 226 working days for that matter / 7,5 hours per day). Most of the people helping out in are between 25 and 44 years of age. March, 2011

4. Digitisation Production – Content Granulation

1. Structural mark up and granulation of a work (aggregates as works)
2. Content granulation/fragmentation in a work (re-use, re-mixing)
3. Automatic extraction of named entities from works (ontologies development, enrichment of authority files)

Structural markup and content mark up as part of digitisation production….

1.Article/Chapter/Illustration (aggregates as works, container and metadata profile, ,digital preservation)
2.Paragraph (annotatin of thema, event, time)
3.Sentence (annotatin of thema, event, time)
4.Word (named entities)
5.Character (OCR correction)



(205)

TRAVELS

THROUGH

FINLAND.

CHAPTER XIII.

Arrival in Finland at the Town of Abo—Account of this City—The Library—The Univerſity—Admiral Heding—Farther Particulars of the Town of Abo; its Situation, Streets, Buildings—The Cathedral—The Building of the Academy—The Harbour—The Inhabitants of Abo—Their Trade and Commerce—The Caſtle, called Abo-hus.

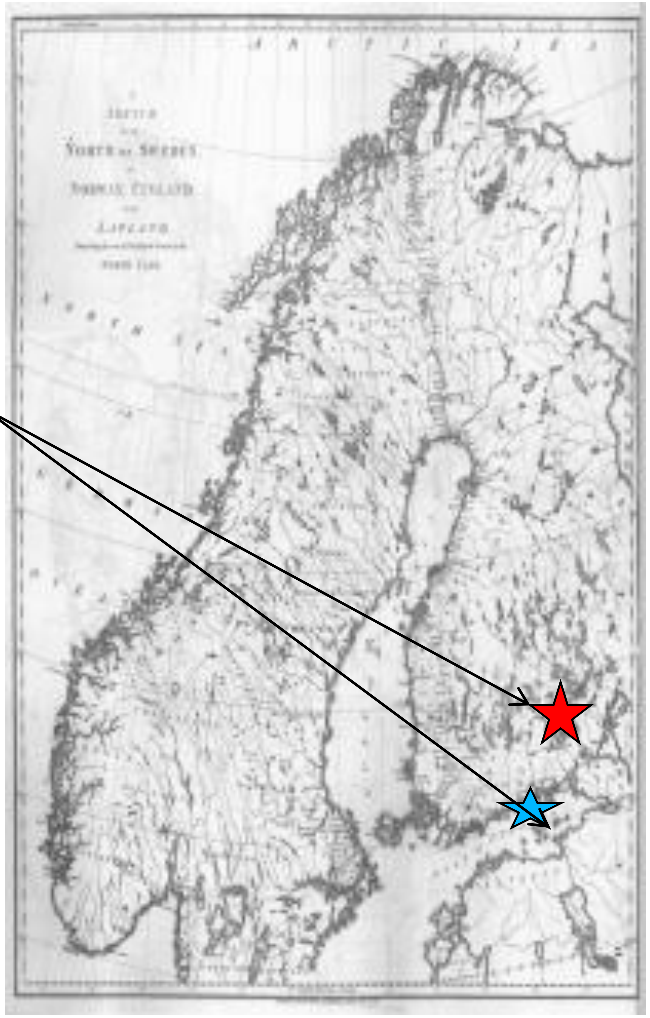
IN our way to Abo we paſſed near the caſtle named Abo-hus, ſituated at the mouth of the river Aura, upon a cape or point of land, bounded by the water on three ſides.* The governor and biſhop, who generally reſide at Abo, happened to be abſent when we arrived; but we had the good fortune to be introduced to Admiral Heding, who received us with great politeneſs, and

engaged

* See a deſcription of it, page 214.

5. Access to Digital Collections….

1. **Physical: National Library of Finland**
2. **Digitisation: Centre for Digitisation and Conservation; full in-house digitisation production chain, 2milj p newspapers, 2 milj p of journals, books, parchments.**
3. **Access to Digitized Collections (Public Domain)**
   http://www.digi.kansalliskirjasto.fi
4. **Restricted Access to Digitized Collections (Copyright Protected)**
5. **Access to Funded Digitisation by Projects**
   http://www.rahasto.kansalliskirjasto.fi/pelastakirja/ or https://www.doria.fi/handle/10024/50699
6. **National Digital Library – Europeana**
   http://www.kdk.fi/en



RDA – 2011 NLF decision to adopt RDA by 2014.

Open linked data on agenda…

Cataloguing Policy due Spring/Summer 2011.

6. Beyond Access – Semantic Contextualization

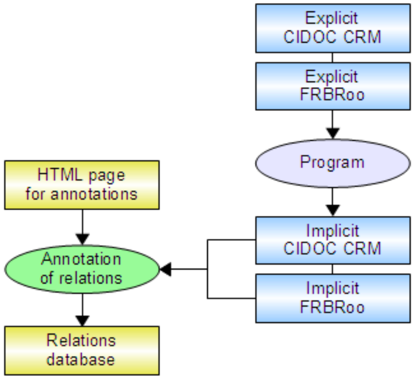
For creating a web of linked resources through digitisation production:

- Containers, Content and Context as part of digitisation production processes as part of distributed digitisation production…
- Focus away from the Catalogue (even though it is magnifique) and make do with a minumum record with persistent ID and link to manifestation 1.
- Containers still need to be specified and maintained for digital manifestation 2 and aggregations of digital manifestation 2. If each work stands as separate entity with its associated metadata, it can be linked and annotated in different contexts (METS profiles for works, and how to update containers…)
- Use of linked and controlled vocabularies and ontologies in digitisation production aid semantic contextualization (YSO, FinOnto, OCM)
- Automatic and semi-automatic extraction of data, named entities, thema, subjects, concpets relations and meaning as part of digitisation production…FRBR, FRAD, FRSAD/CIDOC CRM..
- Integrate distributed work processes and crowd sourcing, in addition to repetitive work such as OCR correction, to structural and content markup….

6. Experimentation with Beyond Access – Annotation/Extraction of Meaning….

Acerbi as prototype:
Using FRBR,; Funcitional Requirements for Bibliographic Record (FRAD and FRSAD) for annotation of Acerbi text for extracting meaning from content– does it work for ?

1. Markup of named entities (FRAD) (VIAF)
2. Markup of time
3. Markup of place (GeoOnto)
4. Markup of thema (OCM)
5. Markup of subject (FRSAD)
6. Markup up of relations between entities
7. Markup of other….Acerbi defined chapter structure, Acerbi's knowledge domain, Acebi's emotions….



How about FRBRoo and CIDOC CRM perspectives ?

How about EDM-FRBRoo perspectives ?

Experimentation; Tiina Ison, Eeva Murtomaa, Mika Nyman - enabling access to content utilizing conceptual models ….

Thank You

tiina.ison@helsinki.fi