Opening the Archives:
OLAC, TRACTOR and the OTA

Martin Wynne
Oxford Text Archive
martin.wynne@ota.ahds.ac.uk

Outline

1. Why did I want to join in with OLAC?
2. What is TRACTOR?
3. What is the Oxford Text Archive?
4. The experiment
5. The results

Standards

- I don't like them;
- But some level of interoperability of metadata is necessary for a centralised resource discovery mechanism;
- Resource encoding standards and metadata standards;
- There are good standards and bad standards.

Bad standards

- Devise elaborate standards in the abstract without reference to practicalities;
- Make complex standards with many obligatory categories;
- Use technologies with which the community is not familiar or which are arcane.

Good standards

- Easy to implement, especially making migration from other formats easy;
- Provide tools and support for migration and implementation;
- Make validation easy;
- Give immediate rewards, e.g. a working search engine;
- Use or relate to existing community (and wider) standards.

Portals

UK:

- Resource Discovery Network
- HUMBUL Humanities Hub
- Arts and Humanities Portal

International:

- Internet search engines
- Subject-based meta-archives (e.g. OLAC)

An alternative: IMDI

- More detailed metadata set
- DC is only the lowest common denominator: need more detailed set for describing language resources properly
- But what categories for what sort of resources?
- IMDI really for linguistic field data, not corpora
- Can now be migrated to OLAC

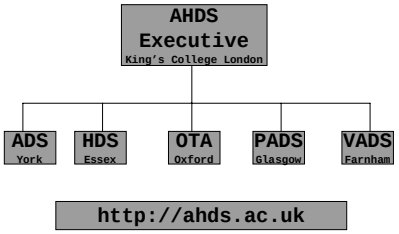
Conclusions so far

- Need a meta-archive so that resources in different archives can more easily be discovered
- OLAC seems to offer a potentially viable model
- Let's try it out with an archive (or two!)

TRACTOR

- A key part of TELRI II
- www.tractor.de
- Resources in *Bulgarian, Croatian, Czech, Dutch, English, Estonian, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Romanian, Russian, Serbian, Slovak, Slovenian, Swedish, Turkish, Ukrainian, Uzbek.*
- Mostly corpora; some lexicons and tools
- c. 100 resources

Arts and Humanities Data Service structure and services



Oxford Text Archive: Holdings

- 25+ languages, c. 2,500 titles
  - Literary texts
  - Reference works, dictionaries, Bibles
  - Corpora (50+)
- Catalogue (TEI SGML and XML, library cataloguing guidelines followed)
- Full text retrieval in various formats
  - HTML, TEI-Lite, ASCII, RTF, original(?)
- Mirror sites

The experiment: prerequisites

- Structured metadata
- Access to standards and ability to migrate
- Validation tools
- Harvest mechanism

Opening TRACTOR and the OTA to OLAC (1)

Two different archives:

TRACTOR: 99 records, corpora, designed for LE, not much metadata

OTA: 2500 records, mostly literature, some corpora, metadata well-structured but partial

Common points: variable quality of resources and metadata

Process

- Sed, awk, commands and scripts and perl programs to migrate to OLAC-conformant metadata records
- Hack XML file of records
- Validate records
- Send XML file to OLAC

Conversion of TRACTOR metadata (1)

- Parse (reasonably well-structured) HTML file to get metadata categories;
- Re-order (in Excel) and edit to fit OLAC categories (and export to tab-separated text);
- Hack OLAC-conformant XML file from the text.

Conversion of TRACTOR metadata (2)

```
<hr>
<h2>
<a NAME="Bulgarian"></a><i>Bulgarian</i></h2>

<ul>
<p><li><h3>
POS tagged corpus</h3>
2460 Bulgarian sentences marked-up with part of speech
   information (BTB-POS Corpus I). The corpus
is in XML format, non-standard with respect to TEI or CES, DTD
   is included. Available in three
different encodings of cyrillic letters: ISO 8879:1986, MS
   Windows, and Unicode.
<p><i>Resource provider:</i> <a
   href="mailto:kivs@bgcict.acad.bg">Kiril Iv. Simov</a>,
   Linguistic Modelling Laboratory, Bulgarian
Academy of Sciences, Sofia.
<p><a href="/tractor/resources/SOF2/BTB-POS/">Browse the
   files</a>
```


Conversion of TRACTOR metadata (3)

```
SOF2001 2001-10-31      "Kiril Iv. Simov, Linguistic Modelling Laboratory, Bulgarian"   2460
    sentences  unknown 2001    "2460 Bulgarian sentences mar
ked-up with part of speech information (BTB-POS Corpus I). The corpus is in XML format, non-
    standard with respect to TEI or CES, DTD is included. Ava
ilable in three different encodings of cyrillic letters: ISO 8879:1986, MS Windows, and Unicode. "
    text            ISO 8879:1986   XML with DTD
                    http://tractor.bham.ac.uk/tractor/resources/SOF2/BTB-POS/       Bulgarian
    www.tractor.de          TRACTOR User Agreemen
t                       BTB-POS Corpus I        corpus
SOF2002 2001-10-31      "Kiril Iv. Simov, Linguistic Modelling Laboratory, Bulgarian"   2460
    sentences  unknown 2001    "2460 Bulgarian sentences mar
ked-up with part of speech information (BTB-POS Corpus I). The corpus is in XML format, non-
    standard with respect to TEI or CES, DTD is included. Ava
ilable in three different encodings of cyrillic letters: ISO 8879:1986, MS Windows, and Unicode. "
    text            Windows XML with DTD
                http://tractor.bham.ac.uk/tractor/resources/SOF2/BTB-POS/       Bulgarian
    www.tractor.de          TRACTOR User Agreement
                    BTB-POS Corpus I        corpus
SOF2003 2001-10-31      "Kiril Iv. Simov, Linguistic Modelling Laboratory, Bulgarian"   2460
    sentences  unknown 2001    "2460 Bulgarian sentences mar
ked-up with part of speech information (BTB-POS Corpus I). The corpus is in XML format, non-
    standard with respect to TEI or CES, DTD is included. Ava
ilable in three different encodings of cyrillic letters: ISO 8879:1986, MS Windows, and Unicode. "
    text            Unicode XML with DTD
                http://tractor.bham.ac.uk/tractor/resources/SOF2/BTB-POS/       Bulgarian
    www.tractor.de          TRACTOR User Agreement
                    BTB-POS Corpus I        corpus
```

Conversion of TRACTOR metadata (4)

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE OLAC-Repository SYSTEM "olacrep.dtd">
<OLAC-Repository>
    <identity>
        <repositoryIdentifier>tractortest</repositoryIdentifier>
        <repositoryName>TRACTOR Test Archive</repositoryName>
        <adminEmail>martin.wynne@ota.ahds.ac.uk</adminEmail>
        <collectionDescription>A test set of metadata records from TRACTOR (TELRI Research Archive of Language Resources and
Tools</collectionDescription>
    </identity>
<sets>
<set>
<setSpec></setSpec>
<setName></setName>
</set>
</sets>
<records>
    <record>
    <header>
    <recordId>olac:tractor:SOF2001</recordId>
    <datestamp>2001-10-31</datestamp>
    </header>
    <metadata>
    <olac lang="en">
        <contributor>"Kiril Iv. Simov, Linguistic Modelling Laboratory, Bulgarian"</contributor>
        <coverage>2460 sentences</coverage>
        <creator>unknown</creator>
        <date>2001</date>
        <description>"2460 Bulgarian sentences marked-up with part of speech information (BTB-POS Corpus I). The corpus is in XML
format, non-standard with respect to TEI or CES, DTD is included. Available in three different encodings of cyrillic letters: ISO 8879:1986,
MS Windows, and Unicode. "</description>
        <format>text</format>
```


Conversion of OTA metadata (1)

```
<!DOCTYPE TEI.2 PUBLIC "-//TEI//DTD TEI Lite 1.0//EN" "teilite.dtd">
<TEI.2>
<TEIHEADER TYPE='TSBD(ER)'>
<FILEDESC>
<TITLESTMT>
<TITLE TYPE="main">Eerec [Electronic resource]</TITLE>
<AUTHOR>Chef&eacuten;cten, de Troyes, 12th cent.</AUTHOR>
<RESPSTMT><RESP>creation of machine-readable version</RESP>
<NAME>Lusignan, Serge</NAME>
</RESPSTMT>
</TITLESTMT>
<EXTENT>
<SEG TYPE="designation">Text data</SEG>
<SEG TYPE="size">(1 file : ca. 206 kilobytes)</SEG>
</EXTENT>
<PUBLICATIONSTMT><AUTHORITY ID='LUSSE1'>Deposited by
<NAME TYPE="PERSON">Lusignan, Serge</NAME>
<NAME TYPE="ORG">Inst. d'Etudes Medi&eacute;vaux Universite de Montreal</NAME>
<ADDRESS><ADDRLINE>Inst. d'Etudes Medi&eacute;vaux</ADDRLINE>
<ADDRLINE>Universite de Montreal</ADDRLINE>
<ADDRLINE>CP 6128 Succ A</ADDRLINE>
<ADDRLINE>H3C 3J7 Montreal PQ</ADDRLINE>
<ADDRLINE>Canada</ADDRLINE>
</ADDRESS>
<DATE></DATE>
</AUTHORITY>
<DISTRIBUTOR>
<NAME KEY="ota" TYPE="organisation">Oxford Text Archive</NAME>
<NAME TYPE="place">Oxford</NAME>
<ADDRESS>
<ADDRLINE><NAME KEY="oucs" TYPE="organisation">Oxford University Computing Services</NAME>
</ADDRLINE>
<ADDRLINE>13 Banbury Road</ADDRLINE>
<ADDRLINE>Oxford</ADDRLINE>
```


Conversion of OTA metadata (2)

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE OLAC-Repository SYSTEM "olacrep.dtd">
<OLAC-Repository>
    <identity>
        <repositoryIdentifier>otatest</repositoryIdentifier>
        <repositoryName>Oxford Text Archive</repositoryName>
        <adminEmail>martin.wynne@ota.ahds.ac.uk</adminEmail>
        <collectionDescription>A test set of metadata records from the Oxford Text Archive
(OTA)</collectionDescription>
    </identity>
<sets>
<set>
<setSpec></setSpec>
<setName></setName>
</set>
</sets>
<records>
    <record>
    <header>
        <recordId>olac:ota:0001</recordId>
        <datestamp>2001-11-6</datestamp>
    </header>
    <metadata>
    <olac lang="en">
        <contributor></contributor>
        <coverage>(1 file : ca. 97.6 kilobytes)</coverage>
        <creator></creator>
        <date></date>
        <description></description>
        <format></format>
        <format.cpu></format.cpu>
        <format.encoding></format.encoding>
        <format.markup></format.markup>
```


Validation

- XMetaL (or any XML parser) to check for well-formedness and conformance to the DTD (olacrep.dtd)
- http://www.w3.org/2001/03/webdata/xsv?docAddrs=http%3A%2F%2Fusers.ox.ac.uk%2F~martinw%2Folac%2Fota.xml&warnings=on&keepGoing=on&style=offline to check against XML Schema

Harvest

- Done manually by Steven Bird and Eva Banik from the XML file I supplied;
- In principle can be harvested from a perl script at the archive website which reads the metadata and looks for updates.


Open Language Archives


Open Language Archives


Open Language Archives

Results

- Metadata successfully harvested, and the virtual data provider accesses the records
- Recall imperfect
- Metadata imperfect
- Migration tools imperfect
- But the future is bright!

Acknowledgements

Thanks for their help to:

- Steven Bird (OLAC and LDC) plus Eva Banik and the ARC people (at Old Dominion University) as developers of the OLAC prototype
- TRACTOR people in Birmingham (esp. Everita Milconoka, Andrius Utka, Wolfgang Teubert)
- OTA people (John Leedham, Michael Popham and Alan Morrison)

Archives and Organisations

- http://www.tractor.de
- http://ota.ahds.ac.uk
- http://www.open-archives.org
- http://www.language-archives.org
- http://www.mpi.nl/ISLE
- http://www.wave.ldc.upenn.edu:8082/olac/
- http://www.linguistlist.org

Documentation

- Warner, S, *Exposing and Harvesting Metadata Using the OAI Metadata Harvesting Protocol: A Tutorial*, http://library.cern.ch/HEPLW/4/papers/3/
- Simons & Bird (2001), *OLAC Metadata Set*, http://www.language-archives.org/OLAC/olacms-20011022.html
- Bird & Simons (2001), *The OLAC metadata set and controlled vocabularies. ACL Workshop on sharing tools and resources for research and education*, http://arXiv.org/abs/cs/0105030