TEI in SEE:
Experience and Prospects

Dr. Milena Dobreva, Assoc. Prof.
Digitisation of Scientific Heritage Department

Institute of Mathematics and Informatics - BAS
Sofia, Bulgaria

Outline

- About the region…
- Examples of TEI projects
- A look to the EC priorities (digital preservation of and access to cultural and scientific heritage perspective)
- Some suggestions: how to increase TEI popularity and use

Where do we stay?

SEE: rich in cultural heritage but underrepresented in the digital space

Examples from the Memory of the World Register of UNESCO

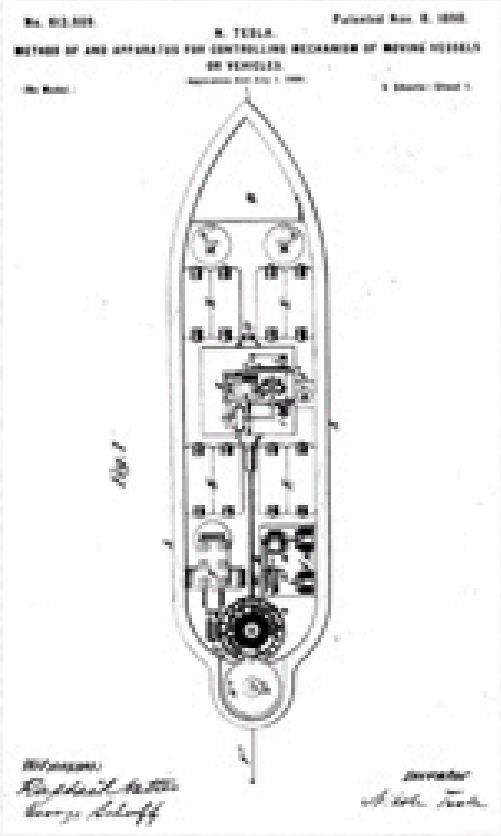
- Albania
  - Codex Purpureus Beratinus (2005)
- Serbia and Montenegro
  - Nikola Tesla's archive (2003)
  - Miroslav Gospel (2005)
- Bulgarian proposal
  - Evangelium Assemanii

Codex Purpureus Beratinus

- Two Gospels, one of 6th Century and of 9th Century
- Only 7 purple codices survive today in Europe

Nikola Tesla's archive

- Collection of manuscripts, photographs, scientific and patent documentation
- Nikola Tesla (1856-1943) - Serbian-born, American inventor and scientist, a pioneer in electrification, significantly influenced the technological development of our civilization by his polyphase system inventions.
- The magnetic induction unit (tesla) of the SI system is named after him.

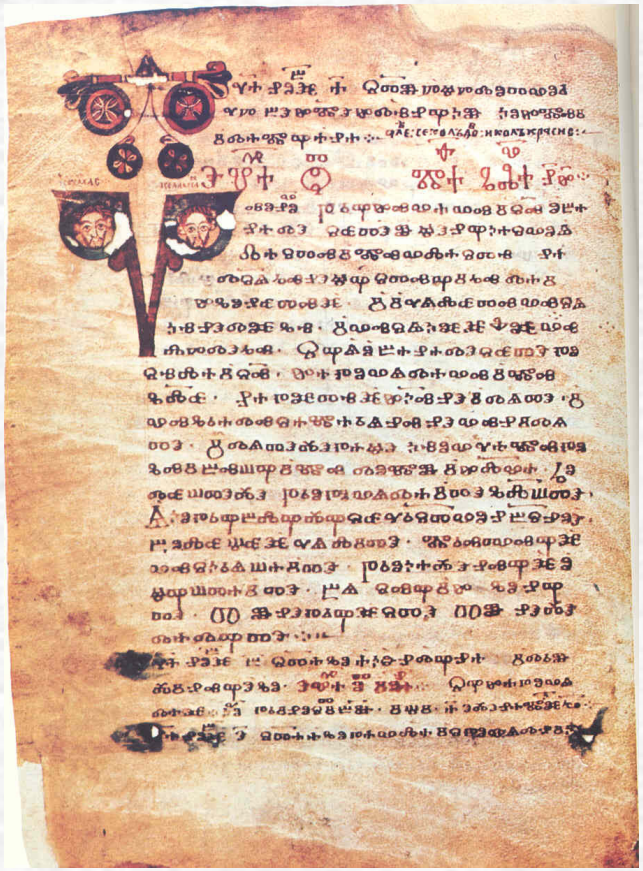


Miroslav Gospel

- Manuscript written in Old Cyrillic from 1180
- Representative of a group of illuminated manuscripts of specific style and iconography resulting from fusion of elements of the West and the East (Byzantium).

Bulgarian proposal

- Evangelium Assemanii,
- Glagolitic script
- End of 10th- beginning of 11th century



Scope example: some Holdings in Bulgarian Repositories

- 8500 Christian manuscripts + 4000 Islamic
- More than 35000 early printed books
- Third largest collection of epigraphic inscriptions in Latin and Ancient Greek in the world
- Proto Bulgarian runic inscriptions
- Immovable cultural heritage monuments
- … and numerous other
  - ***Which have you seen online???***
  - ***Where TEI has been used???***

Lofty or realistic goals?

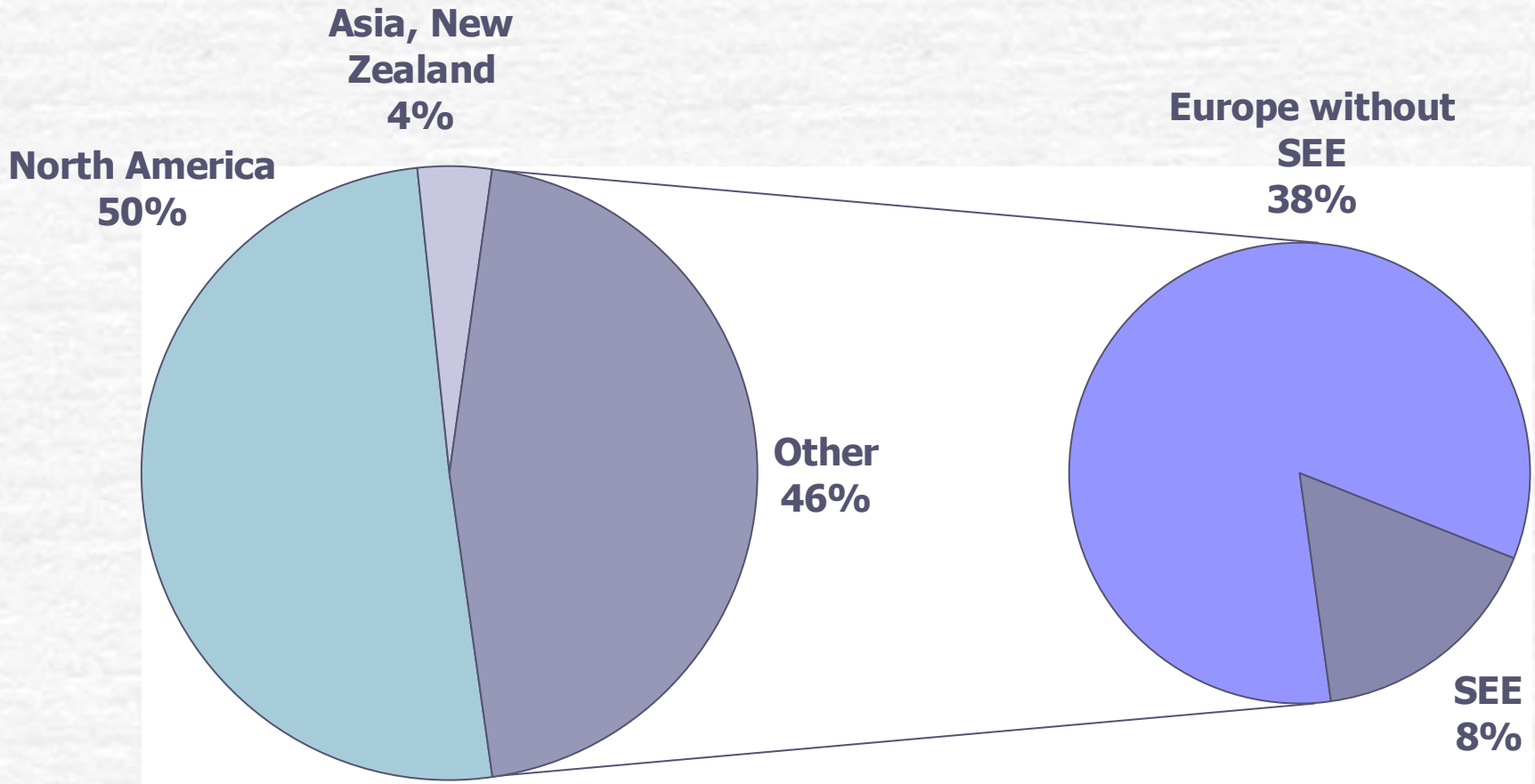
- Content
  - Towards culturally-responsive and high-quality digital content
    - Quality of encoding used
    - Quality of data provided
- Capability
    - Use of existing standards for mark-up and tools
    - Creation of new ones
- Connectivity

The Region

- Albania
- Bosnia and Herzegovina
- **Bulgaria**
- Croatia
- Macedonia (FYROM)
- Moldova
- **Romania**
- **Serbia and Montenegro**
- **Slovenia**

- EC terms: NMS, AC, West Balkan countries, NIS.

Geographic Distribution of TEI members



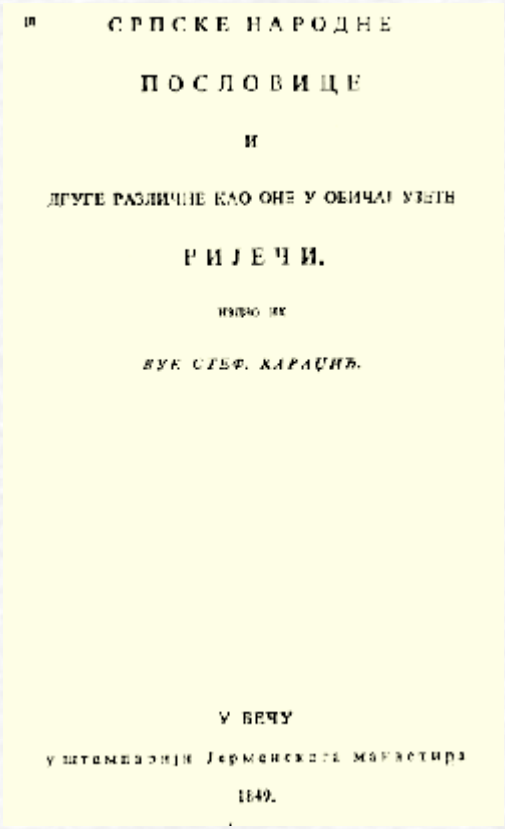
Bulgaria 3, Serbia and Montenegro 1, Slovenia 2

MULTEXT-EAST

- MULTEXT-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages – EC supported project (1996-98)
- Multilingual dataset for language engineering research and development.
- Languages covered initially: **Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene**
- Later additions: **Russian, Serbian, Croatian, Resian, Latvian and Lithuanian**
- Some, or all of the following language resources: the MULTEXT-East morphosyntactic specifications, lexica, and annotated "1984" corpus; the MULTEXT-East parallel and comparable text and speech corpora; and associated documentation.
- All corpora now encoded in XML TEI P4
- URL: http://nl.ijs.si/ME/

The Serbian Proverbs Project

- Based on the Collection of Serbian Proverbs prepared by the reformer of the contemporary Serbo-Croatian language, Vuk Stefanović Karadžić
- Over 6000 proverbs
- New elements defined
- URL: http://alas.matf.bg.ac.yu/~cvetana/proverb/



СРПСКЕ НАРОДНЕ

ПОСЛОВИЦЕ

И

ДРУГЕ РАЗЛИЧНЕ КАО ОНЕ У ОБИЧАЈ УШЕЛЕ

РИЈЕЧИ.

ИЗДАО ИХ

ВУК СТЕФ. КАРАЏИЋ.

У БЕЧУ

У ШТАМПАРИЈИ ЈЕРМЕНСКЕ МАНАСТИРА

1849.

Transpoetika

Digital technologies and distributed learning in foreign language study

- Annotated multimedia editions of the most important works by Serbian and Yugoslav writers
- Four components of the architecture
  - *TransText*: encoded in XML with a TEI DTD with embedded links toward the other modules;
  - *TransDictionary*: a Serbian-English dictionary for the students of Serbian as a foreign language, based on mySQL database;
  - *TransMedia*: a collection of foreign language audio and video materials, saved in the QuickTime format for easy internet delivery; a media component of the TransText and TransDictionary modules (audio recordings of the texts, pronunciation of headwords and examples in the dictionary etc.);
  - *TransForum*: a web-based forum for the discussion of the Serbian language and literature, which can be linked to from the other three modules.
- URL: http://www.transpoetika.org/

Scholarly Digital Editions of Slovenian Literature

- Scientific Research Centre of the Slovenian Academy of Sciences and Arts & Jozef Stefan Institute. Department of Knowledge Technologies. Ljubljana, Slovenia
- Aims at producing a collection of digital critical editions of Slovenian literary texts and sources for literary studies
  - *digitised facsimile*
  - *diplomatic transcription*
  - *critical transcription*
- The digital editions use TEI P4 (XML) for their encoding model, with the following tagsets: prose, figures, linking, and transcription.
- Some local modifications
- URL: http://nl.ijs.si/e-zrc/index-en.html

Repertorium of Old Bulgarian Literature and Letters

- Developed as an "archival repository capable of encoding and preserving in SGML (and, subsequently, XML) format archeographic, palaeographic, codicological, textological, and literary-historical data concerning original and translated medieval texts represented in Balkan Cyrillic manuscripts".
- Started as an initiative of David J. Birnbaum (University of Pittsburgh), Andrej Bojadžiev (University of Sofia), Milena Dobreva (Institute of Mathematics, Bulgarian Academy of Sciences), and Anisava Miltenova (Institute of Literature, Bulgarian Academy of Sciences) in 1994, with early SGML development assistance from Harry Gaylord and Berend Dijk (University of Groningen, The Netherlands).
- Repertorium input to TEI:
  - Developed a manuscript description DTD
  - Contributed to the work of the MASTER project (taken into account within the Taskforce on Manuscript Description at TEI)
- URL: http://clover.slavic.pitt.edu/~repertorium/index.html

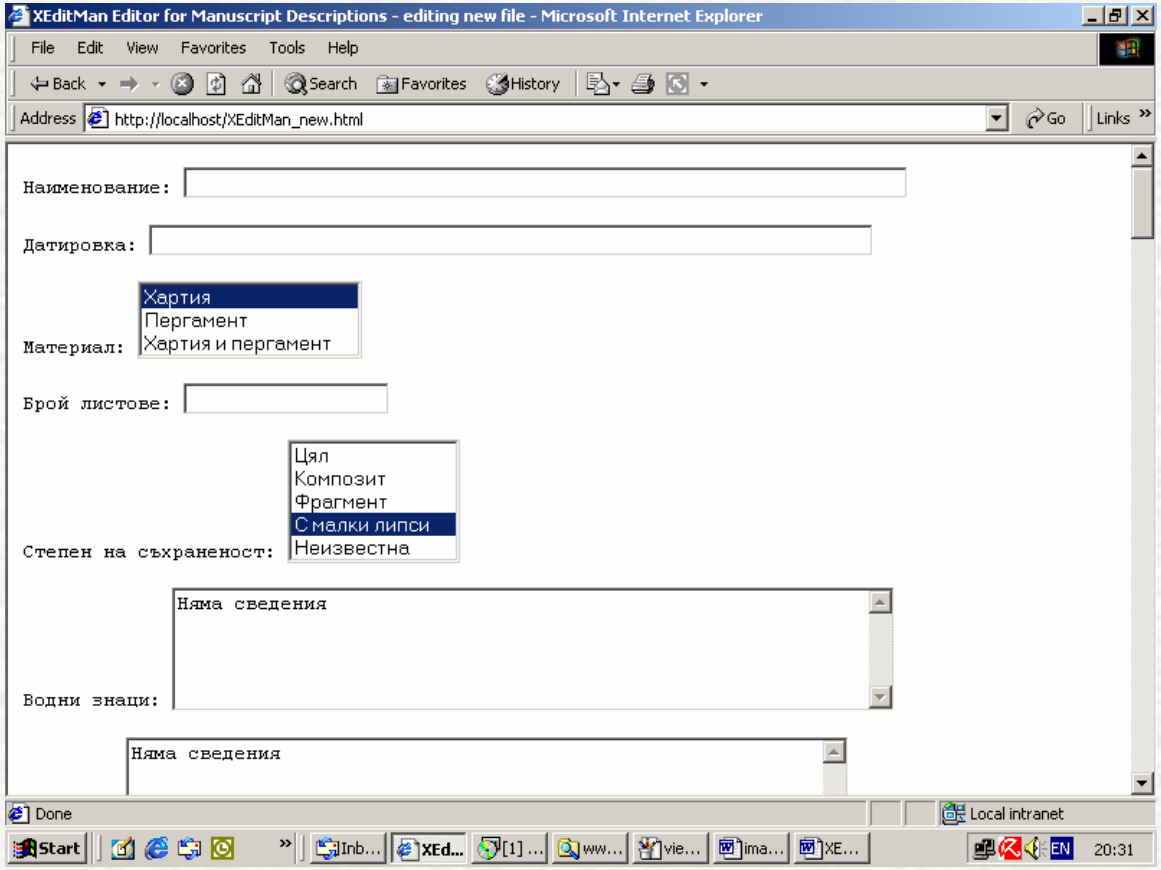
Repertorium follow-up projects

- Computer-Supported Processing of Slavonic Manuscripts and Early Printed Books (Institute of Literature, Bulgarian Academy of Sciences and Central European University, 1997–98)
- Computer Data Base of Late Medieval Bulgarian Literature (Fourteenth–Eighteenth Centuries) (Institute of Literature, Bulgarian Academy of Sciences and Royal Swedish Academy of Letters, History, and Antiquities, 2002–2005)
- Computer Processing and Analysis of Slavic Manuscripts (Institute of Literature, Bulgarian Academy of Sciences and Institute of Russian Literature, Russian Academy of Sciences, 2002–2005)
- Machine-Readable Description and Searchable Catalogues of Cyrillic Manuscripts (Institute of Literature, Bulgarian Academy of Sciences, Central Library, Bulgarian Academy of Sciences, British Library, 2003–2006)

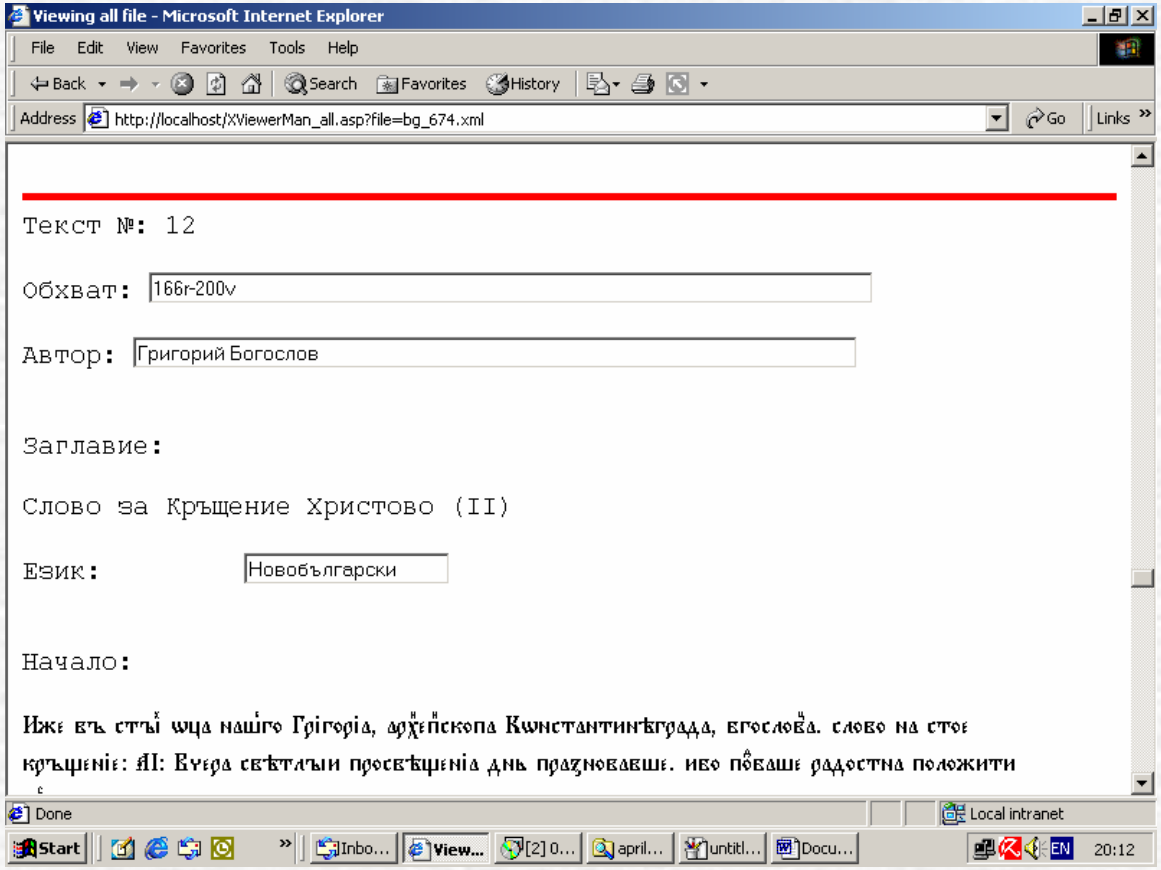
Work on manuscript cataloguing within the KT-DigiCult-Bg project

- XEditMan, XML Editor for Manuscripts, was developed
- It is an integrated tools which offers support for the following activities:
  - Editing a new catalogue description
  - Editing an existing catalogue description
  - Visualisation of one complete description
  - Visualisation of user-selected data from a description
- Executing queries over a group of catalogue descriptions
- Over 800 manuscript descriptions in Bulgarian entered in 2004-2005, XML TEI P4 conformant
- Publication pending (National library)
- Work continues into the direction of intelligent search tools

XEditMan: data entry



XEditMan: visualisation



Popularizing TEI in the region and beyond

- 1995: First International Conference Computer Processing of Mediaeval Slavic Texts (Blagoevgrad)
- 1996: Workshop *Text Variety in Mediaeval Manuscripts*, supported by UNESCO (Sofia)
- 2001: Concluding workshop of the MASTER project supported by the EC (Sofia)
- 2002: Electronic Description and Edition of Slavic Sources (Pomorie)
- 2003: Mini-symposium "Digital Preservation of Cultural Heritage" within the frameworks of the International congress of the MASSEE - Mathematical Society of South-Eastern Europe (Borovets)
- 2003: Computer-Supported Processing of Medieval Slavonic Manuscripts and Early Printed Books (Roundtable). Thirteenth International Congress of Slavists. August–September 2003 (Ljubljana, Slovenia)
- Since 2002, Conferences of the NCD – Serbia and Montenegro
- 2005: First SEEDI conference (Ohrid)

Popularizing TEI in the region and beyond

- Training of specialists from Central/Eastern Europe (international summer schools including TEI mini-courses delivered by Matthew Driscoll and Lou Burnard)
  - *CIDOC conference in Zagreb, May 2005*
  - *Electronic Publishing for Cultural Heritage Studies* supported by the European Commission and the National Agency for ICT (2002)
  - *Digital Preservation of Medieval Manuscripts and Early Printed Books* supported by Open Society Institute, Open Society Foundation-Sofia, IREX-Washington, etc. (1999)
  - *Applications of IT to Biblical Studies* supported by Open Society Institute and Open Society Foundation-Sofia (1998)

International Bodies

- Commission to the Executive Council of the International Committee of Slavists for the Computer-Supported Processing of Slavic Manuscripts and Early Printed Books
- SEEDI (South-East European Digitisation Initiative)

Borovets declaration and SEEDI

- Borovets declaration signed at MASSEE congress in 2003
  - Institute for Mathematics and Informatics, BAS, Bulgaria
  - Institute for Bulgarian Language, BAS, Bulgaria
  - Institute of Information Technologies, BAS, Bulgaria
  - Sofia University, Bulgaria
  - Arnamagnæan Institute, Copenhagen University, Denmark
  - University of Dublin, Trinity College, Dublin, Ireland
  - Unitarian Archives and Library in Cluj/Kolozsvár, Romania
  - Faculty of Mathematics Belgrade University, Serbia and Montenegro
  - Mathematical Institute SANU, Serbia and Montenegro
  - ARISTOS company, Ukraine

Borovets declaration and SEEDI

- Aims – to build a network
  - To mobilise the human and material resources existing in the region;
  - To disseminate scientific information as well as the results of research;
  - To facilitate communication between centres having similar scientific interest.

Input of regional projects to TEI

Project	Adding new elements to TEI	Development of tools
MULTEXT-East	✓	
Serbian proverbs	✓	
Repertorium	✓	✓
XEditMan	no	✓

Timeline

1987	TEI founded		
1994	TEI P3	**1994-95**	Repertorium project
		1995	First international conference in Bulgaria
		1996-98	MULTEXT EAST project Participation in MASTER project
2001	Reestablished as a membership consortium		
2002	TEI P4	**2002-…**	Serbian proverbs, Transpoetika, Scholarly Digital Editions of Slovenian Literature, Slovenian-English corpus, Encoding of Colloquial Bulgarian, Repertorium follow-ups, KT-DigiCult-BG
		2003	SEEDI established

Comparison of the local situation with EC priorities

- Expressed in *Lund Principles*:
  - *making visible and accessible the digitised cultural and scientific heritage of Europe*;
  - *coordination of efforts*;
  - *development of a European view on policies and programmes*, as well as of *mechanisms to promote good practice in a consistent manner*
- EU national libraries own over 100 million items, less then 2% are digitised
- Digitization itself is not supported for several years
- EC recommendations on digitisation are expected in June 2006 (Digital library programme)

DigiCult: TWRs on new/core/emerging technologies in the cultural/scientific heritage sector

- **DigiCULT TWR 3, December 2004**
  - Open Source Software and Standards
  - Natural Language Processing
  - Information Retrieval
  - Location-Based Systems
  - Visualisation of Data
  - Telepresence, Haptics, Robotics;
- **DigiCULT TWR 2, February 2004**
  - The Application Service Model
  - The XML Family of Technologies
  - Cultural Agents and Avatars, Electronic Programming Guides and Personalisation
  - Mobile Access to Cultural Information Resources
  - Rights Management and Payment Technologies
  - Collaborative Mechanisms and Technologies
- **DigiCULT TWR 1, February 2003**
  - Customer Relationship Management
  - Digital Asset Management Systems
  - Smart Labels and Smart Tags
  - Virtual Reality and Display Technologies
  - Human Interfaces
  - Games Technologies

MINERVAplus

Ministerial NEtwoRk for Valorising Activising in digitisation

www.minervaeurope.org

- 114 good practices references
- Recommendations for the quality of cultural institutions' websites

Local specific features

- So far: the highest interest is to the use of TEI for historical documents and parallel corpora
- European importance of collections…
  …but they are not accessible in electronic form
- Main experience in pre-digitisation work, such as cataloguing, and text encoding…
  … but mass digitisation never started anywhere
- Digitisation work per se has not been done…
  … thus we do not match current EC priorities
- No governmental programme (resp., funding)…
  … i.e. external financial support is needed
- Regional cooperation is realistic

Conclusions: SWOT analysis STRENGTHS

- Experience already available.
  - Encoding
  - Additions
  - Tools
- Active dissemination.
- Trainings done on a regular basis.
- Work in this region popularizes TEI in wider communities.
- Established professional bodies.

Conclusions: SWOT WEAKNESES

- TEI mostly used in several endeavors of just several key players.
- TEI used for quite specific purposes.
- Scattered experience.
- No mass digitization where TEI could be used extensively (metadata, text encoding).
- Strong dependence on external funding.

Conclusions: SWOT OPPORTUNITIES

- Numerous possibilities - new beginnings could be directed to TEI.
- Work in multilingual setting.
- In the future mass digitization projects TEI could be used more actively (metadata, text encoding).
- Local specifics may provide interesting cases.

Conclusions: SWOT THREATS

- Copyright issues.
- Small projects.
- Various level of relevant experience.
- Lack of crosswalks.
- Promoting TEI means work in conditions where neither governmental nor institutional policies are well established.

Conclusions: increase TEI use

- Connection to more ongoing projects
  - Expanding/improving those which are TEI-related
  - Introducing TEI to those which still do not use it
- Special attention to bodies defining standards where TEI could be applied (preservation and access to cultural and scientific heritage).
- Joint work with repositories planning to start digitisation activities.
- More in-depth exchanges with local specialists.

Conclusions – increase the TEI use

- Introducing TEI to relevant university programmes (LIS, humanities, computer science)
- Dealing with local specific features (character encodings, conceptual)
- More educational work on available tools and "how to" approaches

So… what's next?

Thank you for your attention!