Metadata Interoperability

Lina Bountouri

Ionian University

February 2008

Διαδίκτυο

- Κύριος φορέας πληροφόρησης
- Όγκος δεδομένων
  - **2005**: The "Indexable" Web is 19.1 billion documents [Yahoo, 2005]
- Ποικίλες ομάδες χρηστών
- Ποικίλες ανάγκες πληροφόρησης

- **Αποτέλεσμα: Αναγκαία και παράλληλα δύσκολη η δυνατότητα αποτελεσματικής πρόσβασης**

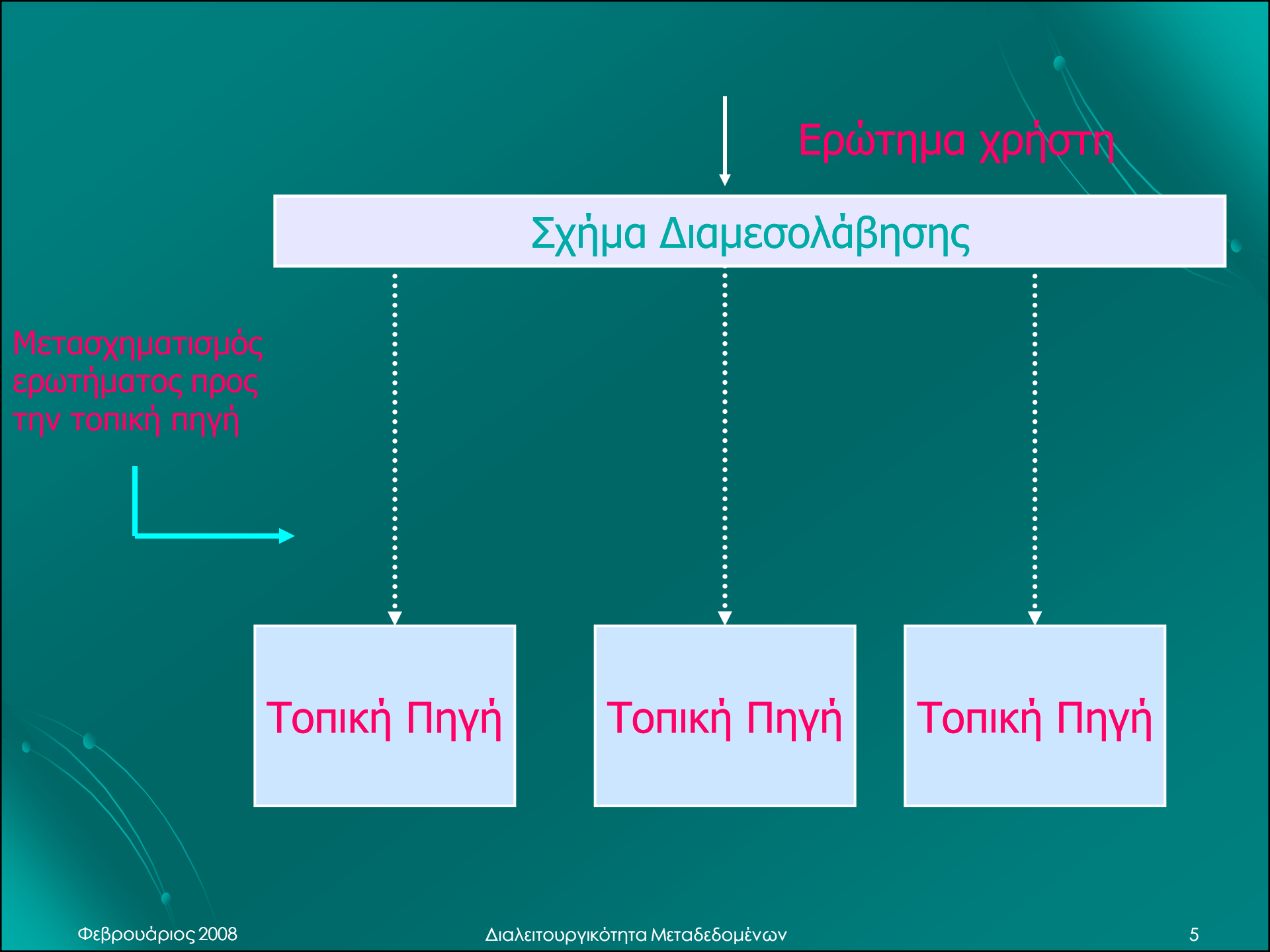
Επίπεδα ετερογένειας δεδομένων

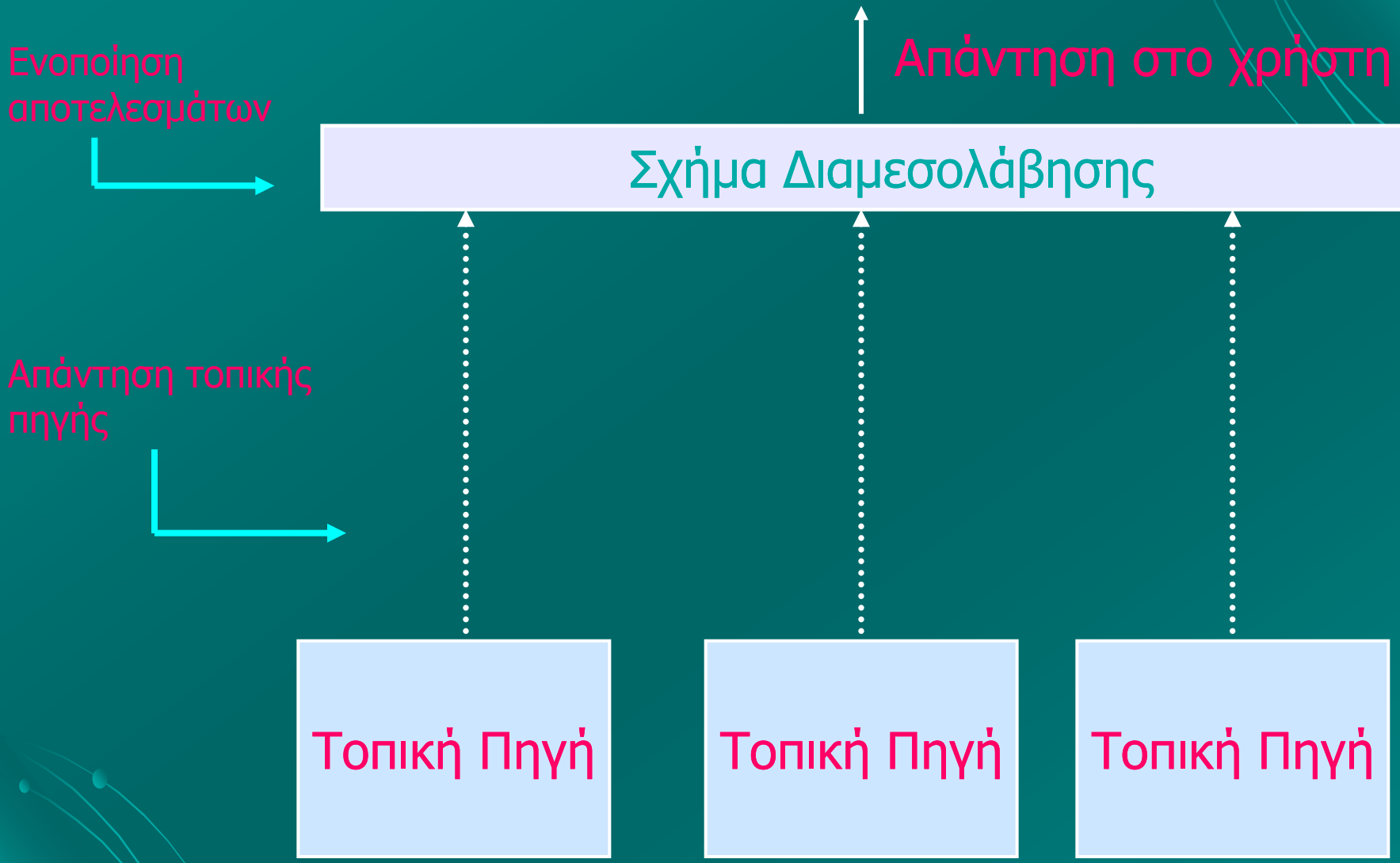
- Ετερογένεια Συστημάτων (hardware, operating systems, networking protocols κτλ)
- Ετερογένεια στη Σύνταξη (query languages, encodings etc)
- Ετερογένεια Σχημάτων (data models, data schemas etc)
- Σημασιολογική Ετερογένεια (semantic conflicts…)

Προβλήματα στην πρόσβαση

Ολοκλήρωση δεδομένων

- Ενιαία πρόσβαση σε συλλογές αυτόνομων πηγών ως ενιαίο σύνολο
- Ένα σύστημα ολοκλήρωσης δεδομένων προωθεί την πρόσβαση και την αποτελεσματική ανάκτηση, εφόσον
  - συνδυάζει δεδομένα από διαφορετικές πηγές, διαφορετικά συστήματα, τα οποία είναι μοντελοποιημένα με διαφορετικά σχήματα
  - παρέχει στους χρήστες ένα ομογενοποιημένο σύνολο αποτελεσμάτων
- Παράδειγμα: Συστήματα Διαμεσολάβησης (Mediated Systems)
  - τα δεδομένα ολοκληρώνονται μέσα από μία εικονική καθολική όψη των επιμέρους πηγών (το καθολικό σχήμα ή σχήμα διαμεσολάβησης (mediated schema or global schema))
  - ο χρήστης ρωτάει το καθολικό σχήμα και το ερώτημα μετασχηματίζεται σε ερώτημα αναγνωρίσιμο από κάθε τοπική πηγή (reformulation step)


Ερώτημα χρήστη
Σχήμα Διαμεσολάβησης
Μετασχηματισμός ερωτήματος προς την τοπική πηγή
Τοπική Πηγή
Τοπική Πηγή
Τοπική Πηγή


Ενοποίηση αποτελεσμάτων
Απάντηση στο χρήστη
Σχήμα Διαμεσολάβησης
Απάντηση τοπικής πηγής
Τοπική Πηγή
Τοπική Πηγή
Τοπική Πηγή

Ολοκλήρωση δεδομένων

- Οι χρήστες ενδιαφέρονται για ενοποιημένες διαδρομές
  - Ολοκλήρωση δεδομένων: (a) δε χρειάζεται να γνωρίζουν που βρίσκονται τα πραγματικά δεδομένα, και (b) να έχουν πρόσβαση σε κάθε μία πηγή χωριστά για να τα ανακτήσουν
- Οι προκλήσεις ενός συστήματος ολοκλήρωσης δεδομένων αφορούν
  - Το βαθμό αυτοματοποίησης της διαδικασίας της ολοκλήρωσης
  - **Τη διατήρηση της σημασιολογίας κάθε πηγής**

Σημασιολογική Ολοκλήρωση Δεδομένων

- Παγκόσμιος Ιστός ⟶ Σημασιολογικός Ιστός
- Ορισμός: *Διαδικασία της χρήσης εννοιολογικών αναπαραστάσεων των δεδομένων και των σχέσεών τους με στόχο της εξάλειψη της σημασιολογικής ετερογένειας*
- Νέες τεχνολογίες για την αναπαράσταση γνώσης και την αξιοποίησή της, π.χ. οντολογίες

Οντολογία

Μία τυπική (formal), κατηγορηματική (explicit) προδιαγραφή μίας διαμοιρασμένης (shared) εννοιολογικής αναπαράστασης (conceptualization)

"Artificial-intelligence and Web researchers have co-opted the term for their own jargon, and for them an ontology is a document or file that formally defines the relations among terms. The most typical kind of ontology for the Web has a taxonomy and a set of inference rules."

Οι Οντολογίες στη Σημασιολογική Ολοκλήρωση Δεδομένων

- Μηχανισμός διαλειτουργικότητας
- Οντολογία ως σχήμα διαμεσολάβησης στην ολοκλήρωση δεδομένων (Ontology-based Integration scenarios)
  - Πολύπλοκες σχέσεις ενός θεματικού πεδίου
  - Αυστηρό φορμαλισμό ⟶ εξαγωγή συμπερασμάτων *(reasoning)*
  - Ανάπτυξη οντολογιών για ένα θεματικό χώρο, εννοιολογική αναπαράσταση ενός τομέα (“*Domain Ontologies*”), π.χ. CIDOC και ABC για πολιτιστικό περιεχόμενο,
  - Ευθυγράμμιση (alignment) εννοιών και όρων

Μεταδεδομένα

"The word is half Latin and half Greek. No good can come of it!"

Τα μεταδεδομένα είναι οι πληροφορίες προσδιορίζουν και ταυτίζουν πηγές / αντικείμενα, υπάρχουν σε διάφορες μορφές και στόχος τους είναι να καλύψουν ποικίλα επίπεδα περιγραφής, ανάγκες τεκμηρίωσης και πρόσβασης

Real World Example…

- Οδηγώ στο δρόμο για το σπίτι…
- Ψάχνω για φούρνο…
- Βλέπω ταμπέλες καταστημάτων…
- Διακρίνω μία ταμπέλα που γράφει ***«Φούρνος Βενέτης»***
- Παρκάρω και πάω να ψωνίσω…
- Είναι νωρίς το βράδυ (09:00 μ.μ.)…
- Είναι κλειστό…

Real World Example…

- Θα με είχε διευκολύνει αν η ταμπέλα έγγραφε ***«Λειτουργεί ώρες καταστημάτων»***…
- Συμπέρασμα 1: τα μεταδεδομένα είναι παντού και μας βοηθούν να βρούμε την πληροφορία που χρειαζόμαστε!
- Συμπέρασμα 2: δεν είναι πάντα πλήρη ως προς τις ανάγκες πληροφόρησης που καλύπτουν!
- Σκέψη: ένα πληροφοριακό σύστημα δε διαβάζει ταμπέλες, διαβάζει όμως δομημένη, μηχανογραφημένη πληροφορία. Για παράδειγμα, μέσα από ένα δικτυακό τόπο που θα παρείχε αναζήτηση για τα καταστήματα εντός Αθηνών και τα ωράριά τους, θα μπορούσαμε να είχαμε πάρει τις πληροφορίες

Μεταδεδομένα

- Για ένα word file στο PC: Τίτλος αρχείου, δημιουργός, τελευταία τροποποίηση, ημερομηνία δημιουργίας, τοποθεσία στο δίσκο κτλ

Πιο κοντινά παραδείγματα στον επιστήμονα της πληροφόρησης:

- Για ένα CD: Τίτλος, συνθέτης, τραγούδια, τραγουδιστής, στιχουργός, μουσικός παραγωγός, θέση στο ράφι ή/και URL, είδος μουσικής κτλ
- Για ένα έγγραφο του δημόσιου τομέα: τμήμα / διεύθυνση κτλ που το παρήγαγε, αριθμός πρωτοκόλλου, υπογραφές υπευθύνων, νόμος – διάταγμα που ορίζει τη δημιουργία του, απευθυνόμενο κοινό / υπηρεσία κτλ

Μεταδεδομένα

- Περιγραφή και τεκμηρίωση ποικίλλων τύπων υλικού: δημιουργία και χρήση συγκεκριμένων σχημάτων περιγραφής σε διεθνές και εθνικό επίπεδο
  - GILS, GovML, AGLS, eGMS κτλ για κυβερνητική πληροφορία (ενεργά αρχεία)
  - Encoded Archival Description για τα ανενεργά (και ημί-ενεργά) αρχεία
  - Διάφορα MARC για ποικίλου τύπου υλικό που διατίθεται από βιβλιοθηκονομικά συστήματα
  - VRA για εικόνες κυρίως

Μεταδεδομένα

- Ανάγκη σε ενοποιημένη πρόσβαση σε ετερογενές υλικό
  - Για παράδειγμα: υλικό πολιτιστικών ιδρυμάτων (Αρχεία, Βιβλιοθήκες και Μουσεία) [βλέπε τα διάφορα έργα που έχει χρηματοδοτήσει η Κοινωνία της Πληροφορίας με σαφείς οδηγίες υλοποίησης για **«ανοιχτό, διαλειτουργικό περιεχόμενο»**]

Διαλειτουργικότητα Μεταδεδομένων

- Η συμβατότητα ανάμεσα σε δύο ή περισσότερα σχήματα μεταδεδομένων
- Με πρακτικούς όρους, η διαλειτουργικότητα μεταδεδομένων αντικατοπτρίζει την ικανότητα ενός συστήματος να συσχετίζει την εννοιολογική (και όχι μόνο) πλευρά ενός σχήματος μεταδεδομένων με ένα άλλο σχήμα

Διαλειτουργικότητα Μεταδεδομένων

- Τα μεταδεδομένα υλοποιούνται μερικές φορές με τρόπο που δεν τα καθιστά διαλειτουργικά
  - Μη σαφείς ορισμοί της σημασιολογίας των πεδίων
  - Λανθασμένη χρήση πεδίων
  - **Βασικό Πρόβλημα: "There are nearly as many types of metadata as there are digital collections!!!"**
    - Application Profiles (☹)…

Διαλειτουργικότητα Μεταδεδομένων

- **Metadata element (πεδία μεταδεδομένων)**: Ορίζει μία αρκετά abstract αλλά παρόλα αυτά σαφή έννοια για το χαρακτηρισμό δεδομένων. Για παράδειγμα, το πεδίο "Creator" στο DC ορίζει την πρώτη πνευματική υπευθυνότητα για τη δημιουργία του περιεχομένου
- **Metadata element instance (στιγμιότυπα των πεδίων των μεταδεδομένων)**: Ορίζει ένα συγκεκριμένο σύνολο δεδομένων μέσα στο πεδίο και σύμφωνα με την έννοια του περιεχομένου του πεδίου (metadata element). Για παράδειγμα, το instance του πεδίου ‘‘Creator’’ για αυτή την παρουσίαση είναι Λίνα Μπουντούρη

Διαλειτουργικότητα Μεταδεδομένων

- **Metadata Schema (σχήμα μεταδεδομένων)**: ορίζει ένα σύνολο σαφών πεδίων μεταδεδομένων (σαν αυτά που αναφέραμε EAD, DC, GILS κτλ)
- **Metadata schema instance (στιγμιότυπα του σχήματος των μεταδεδομένων)**: Ορίζει ένα συγκεκριμένο σύνολο δεδομένων που είναι σύμφωνα με το σχήμα μεταδεδομένων και των εννοιών των πεδίων του. Για παράδειγμα, όλα τα μεταδεδομένα για αυτή την παρουσίαση περιγράφονται με τα instances ενός σχήματος, π.χ. του DC

Μηχανισμοί διαλειτουργικότητας

- Σε γενικές γραμμές δύο βασικές προσεγγίσεις
  - Ορισμός ενός κοινού προτύπου (όπως έχει γίνει στο χώρο των βιβλιοθηκών με τα διάφορα MARC **(;)**)
  - Δημιουργία "*metadata gateways*" οι οποίες μετατρέπουν συγκεκριμένα σχήματα μεταδεδομένων σε άλλα σχήματα
- Έμφαση στη δεύτερη προσέγγιση

Μηχανισμοί διαλειτουργικότητας

- Στα πλαίσια της δεύτερης προσέγγισης αξιοποιούνται διάφορα εργαλεία, όπως:
  - Interoperable core: ένα σχήμα μεταδεδομένων (σχετικά απλό, συνήθως το DC) αξιοποιείται ως κεντρικό σημείο αναφοράς και γίνονται συσχετίσεις από τα διάφορα μεταδεδομένα προς αυτό
  - Crosswalks: ορισμός συσχετίσεων των πεδίων ενός σχήματος μεταδεδομένων προς ένα άλλο
  - Ontology-based integration: Ολοκλήρωση δεδομένων με τη χρήση οντολογιών, όπου μία οντολογία αξιοποιείται ως κεντρικό σημείο αναφοράς και γίνονται συσχετίσεις από τα διάφορα μεταδεδομένα προς αυτήν

Τι εξετάζουμε εμείς;

- Crosswalks από αρχειακά μεταδεδομένα (π.χ. EAD) προς βιβλιογραφικά μεταδεδομένα (π.χ. MODS)
- Συσχετίσεις από μεταδεδομένα που αφορούν το χώρο των "*cultural/memory institutions*" προς την οντολογία τεκμηρίωσης πολιτισμικής πληροφορίας CIDOC CRM

Crosswalks

- Ένα πεδίο ενός σχήματος να αντιστοιχεί σε ένα ή περισσότερα πεδία ενός άλλου σχήματος
  - Για παράδειγμα το πεδίο did/unittitle του EAD αντιστοιχεί στο πεδίο titleInfo/title του MODS
  - Ιδανικό για περιορισμένο αριθμό σχημάτων μεταδεδομένων και κυρίως για κάλυψη «τοπικών» αναγκών (π.χ. Μετατροπή σε MARC για πρόσβαση μέσω του OPAC)
- Με διάφορα εργαλεία, π.χ. MarcEdit και χρήση stylesheets

Crosswalks

- Ορισμός ποικίλων σχέσεων μέσα από μία συσχέτιση
  - 1: Πολλά
  - 1: Κανένα
  - Δομή – Ιεραρχία
  - Δομή – Σειρά
- Μέθοδος περιορισμένης αποτελεσματικότητας, ειδικά όταν αυξάνονται τα εμπλεκόμενα μεταδεδομένα

Συσχετίσεις μεταδεδομένων με οντολογίες

- Ενδεικτικό παράδειγμα
  - Ορίζουμε το CIDOC ως οντολογίας διαμεσολάβησης σε ένα σενάριο σημασιολογικής ολοκλήρωσης, με την οποία συσχετίζονται διάφορα σχήματα μεταδεδομένων
  - EAD προς CIDOC
    - Encoded Archival Description: Είναι το διεθνές σχήμα μεταδεδομένων για την περιγραφή αρχείων, το οποίο διατηρεί την ιεραρχία του αρχείου και δείχνει το περιεχόμενο των περιγραφικών οδηγών των αρχείων (εργαλεία έρευνας)
    - CIDOC Conceptual Reference Model (CRM): παρέχει ορισμούς και δομή για την περιγραφή των («ορισμένων» και «υπονοούμενων») εννοιών και σχέσεων που χρησιμοποιούνται στην τεκμηρίωση της πολιτιστικής κληρονομιάς

Συσχετίσεις μεταδεδομένων με οντολογίες

- Γιατί χρησιμοποιήθηκε το CIDOC ως μεσολαβητής των μεταδεδομένων;
  - Μοναδική οντολογία για πληροφορία που προέρχεται από αρχεία, βιβλιοθήκες και μουσεία μέχρι στιγμής σε διεθνές επίπεδο. ***(Προτροπή: να ακολουθούμε διεθνή πρότυπα και οδηγίες υλοποίησης. Ακόμα και αν δε μας καλύπτουν πλήρως, μας παρέχουν ως ένα βαθμό συμβατότητα διότι χρησιμοποιούνται και από άλλους!)***

Συσχετίσεις μεταδεδομένων με οντολογίες– Μέθοδος και προβλήματα

- **Path-Oriented Approach**
  - Mappings: Μία συσχέτιση από το σχήμα πηγής (π.χ. EAD) προς το σχήμα στόχος (π.χ. CIDOC) μετατρέπει πεδία από το πρώτο σχήμα στα έγκυρα αντίστοιχα πεδία του σχήματος στόχου
  - Μεθοδολογία για mappings: Μετάφραση των διαδρομών των μεταδεδομένων προς τα σημασιολογικά αντίστοιχα CIDOC μονοπάτια

Συσχετίσεις μεταδεδομένων με οντολογίες– Μέθοδος και προβλήματα

- Ένα μονοπάτι CIDOC έχει τη μορφή entity-*property*-entity, για παράδειγμα
  - E22 (Man Made Object)-*P108(has produced/was produced by)*- E12 (Production Event)-*P14(carried out by/performed)*-E39(Actor) δηλώνοντας ότι ένας δημιουργός (Actor, E39) κατά τη διάρκεια ενός γεγονότος δημιουργίας (E12) δημιούργησε ένα φυσικό αντικείμενο που προέρχεται από ανθρώπινη δραστηριότητα, π.χ. ένα αρχείο (Man Made Object, E22)

Συσχετίσεις μεταδεδομένων με οντολογίες– Μέθοδος και προβλήματα

- Ένα μονοπάτι μεταδεδομένων (κυρίως XML) ορίζεται ως μία ακολουθία από στοιχεία, υπό-στοιχεία κτλ, ξεκινώντας από τον αρχικό κόμβο (στοιχείο) του σχήματος και συνεχίζοντας με τα υπό-στοιχεία που εμπλέκονται κάθε φορά διαχωρισμένα με το σύμβολο (/), για παράδειγμα
  - /ead/archdesc/did/origination/persname, το οποίο δηλώνει το όνομα του δημιουργού της αρχειακής περιγραφής

Συσχετίσεις μεταδεδομένων με οντολογίες– Μέθοδος και προβλήματα

- Η μέθοδος είναι κατάλληλη εφόσον τα μεταδεδομένα και οι οντολογίες κωδικοποιούν τις πληροφορίες μέσω μονοπατιών (paths)
  - Παράδειγμα που αποδεικνύει τη χρήση της μεθόδου: ίδιο στοιχείο (όνομα στοιχείου) σε διαφορετικό μονοπάτι κάθε φορά, με διαφορετική σημασιολογία σε κάθε περίπτωση
    - /ead/archdesc/did/originator/**corpname**
    - /ead/archdesc/did/repository/**corpname**

Συσχετίσεις μεταδεδομένων με οντολογίες– Μέθοδος και προβλήματα

- Μεταδεδομένα και οντολογίες: διαφορετικός σκοπός και λειτουργία
  - Μεταδεδομένα: περιγραφή, ταυτοποίηση, διευκόλυνση πρόσβασης, χρήση και διαχείριση πηγών
  - Οντολογίες: "conceptualization" συγκεκριμένων χώρων και θεμάτων, δεν περιλαμβάνουν πεδία για περιγραφή αλλά έννοιες ενός χώρου και τις σχέσεις μεταξύ τους
  - Διαφορετικός τρόπος ορισμού της σημασιολογίας, οι οντολογίες έχουν δομικά στοιχεία ικανά να εκφράσουν πλούσια σημασιολογία και σχέσεις ανάμεσα στις σημασίες, τα μεταδεδομένα στην παρούσα φάση όχι.

Συσχετίσεις μεταδεδομένων με οντολογίες– Μέθοδος και προβλήματα

- Event orientation
  - Μεταδεδομένα: στοχεύουν στο περιγραφόμενο αντικείμενο (π.χ. αρχείο, άρθρο από περιοδικό). Η οντολογία CIDOC βασίζεται σε γεγονότα και δραστηριότητες (event based).
    - Βασικές έννοιες της οντολογίας είναι τα γεγονότα / δραστηριότητες και η παρουσία άλλων οντοτήτων όπως Actors, Dates, Places, Objects, etc, απαιτεί συχνά τη συμμετοχή τους σε ένα γεγονός ή μία δραστηριότητα
- Wrapper elements
  - Τα περισσότερα XML μεταδεδομένα (π.χ. EAD, TEI και MODS) αποτελούνται από πολλά "wrapper" στοιχεία (π.χ. <did> στο EAD) τα οποία ομαδοποιούν σχετικές πληροφορίες, αλλά δεν έχουν σημασιολογική αξία
  - Δεν τα αξιοποιούμε στο mapping

Συσχετίσεις μεταδεδομένων με οντολογίες– Παράδειγμα

- Ένα EAD έγγραφο αποτελείται από τα μεταδεδομένα του ίδιου του εγγράφου [στοιχείο <eadheader>] και από τα μεταδεδομένα του αρχείου (αρχειακή περιγραφή) [στοιχείο <archdesc>]
- Η αρχειακή περιγραφή μέσω του CIDOC ορίζεται μέσα από τις εξής έννοιες
  - E31 (Document) και E33 (Linguistic Object), δηλώνοντας ότι η αρχειακή περιγραφή είναι ένα κείμενο το οποίο τεκμηριώνει ένα αρχείο
  - E22 (Man-Made Object), δηλώνοντας ότι το αρχείο είναι ένα φυσικό αντικείμενο το οποίο προήλθε από ανθρώπινη δραστηριότητα
  - E73 (Information Object) και E33 (Linguistic Object), όπου οι κλάσεις αναφέρονται σε αντικείμενα τα οποία περιλαμβάνουν την ανθρώπινη μνήμη και δεν εξαρτώνται από κανένα φυσικό μέσο (physical carrier)
  - /ead/archdesc: {E31 (Document), E33(Linguistic Object)}-P70 (documents/is documented in)-E22 (Man-Made Object)-P128 (carries/is carried by)-{E73 (Information Object), E33 (Linguistic Object)}

Βιβλιογραφία

- Christophe Blanchi and Jason Petrone. "Distributed Interoperable Metadata Registry" D-Lib Magazine Vol.7 No. 12 December, 2001. [http://www.dlib.org/dlib/december01/blanchi/12blanchi.html]
- Conrad Taylor. Metadata's many meanings and uses. [http://www.ideography.co.uk/briefings/pdf/PB_metadata.pdf]
- Lorcan Dempsey. *Network Resource Discovery: a European Library Perspective*. In *Libraries, networks and Europe: a European networking study*. Neil Smith (ed). London: British Library Research & Development Department, 1994. [http://www.lub.lu.se/UB2proj/LIS_collection/lorcan.html]
- A. Gulli and A. Signorini. In Proceedings of the WWW 2005, May 10–14, 2005, Chiba, Japan. [http://www.cs.uiowa.edu/~asignori/web-size/size-indexable-web.pdf]

Βιβλιογραφία

- J.D. Ullman. Information Integration Using Logical Views. *Theoretical Computer Science*, 239(2):189-210, 2000.
- Lina Bountouri and Manolis Gergatsoulis. "Interoperability between archival and bibliographic metadata". First International Workshop on Cultural Heritage on the Semantic Web (in conjunction with the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference), 12 November, Busan, Korea, 2007. (Poster)
- Thomais Stasinopoulou, Lina Bountouri, Constantia Kakali, Irene Lourdi, Christos Papatheodorou, Martin Doerr and Manolis Gergatsoulis. "Ontology-based Metadata Integration in the Cultural Heritage Domain". In D.H.-L. Goh, I. Sølvberg, E. Rasmussen, T.H. Cao (eds.), Asian Digital Libraries - Looking Back 10 Years and Forging New Frontiers. 10th International Conference on Asian Digital Libraries, ICADL 2007, Hanoi, Vietnam, December 10-13, 2007, Proceedings. Lecture Notes in Computer Science, Vol. 4822, pages 165-175, Springer-Verlag, 2007.

Βιβλιογραφία

- Constantia Kakali, Irene Lourdi, Thomais Stasinopoulou, Lina Bountouri, Christos Papatheodorou, Martin Doerr and Manolis Gergatsoulis. "Integrating Dublin Core metadata for cultural heritage collections using ontologies". In Proceedings of the International Conference on Dublin Core and Metadata Applications 2007, August 27-31, Singapore, 2007.
- Tim Berners-Lee, James Hendler and Ora Lassila. The Semantic Web. Scientific American Magazine - May, 2001. [http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21]
- M. Lenzerini. Data Integration: A Theoretical Perspective. In Proceedings of the 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS02), June 3-5, Madison, Wisconsin, USA, pages 233-246. ACM, 2002.