# Δεδομένα Βιβλιοθηκών και Δεδομένα Κοινωνικών Επιστημών στον Παγκόσμιο Ιστό

Χρήστος Παπαθεοδώρου (papatheodor@ionio.gr)

Ομάδα Βάσεων Δεδομένων και Πληροφοριακών Συστημάτων,
Τμήμα Αρχειονομίας, Βιβλιοθηκονομίας και Μουσειολογίας,
Ιόνιο Πανεπιστήμιο

και

Μονάδα Ψηφιακής Επιμέλειας,
Ινστιτούτο Πληροφοριακών Συστημάτων και Προσομοίωσης
Ερευνητικό Κέντρο «Αθηνά»

DBIS
database & information systems group
ionian university

# Η αρχή: κωδικοποίηση

- Οι βιβλιοθήκες ανέκαθεν διαχειρίζονταν ετερογενές υλικό
- Οι κανόνες καταλογογράφησης εστίαζαν στο αντικείμενο (object)
- Οι κανόνες καταλογογράφησης εκφράζονταν/υλοποιούνταν από αντίστοιχη κωδικοποίηση που πρόσφερε η διάταξη MARC
- Βασική αρχή η διαλειτουργικότητα
  - Τα βιβλιοθηκονομικά πρότυπα αναπτύχθηκαν με σκοπό τη διαλειτουργικότητα και την επικοινωνία μεταξύ μεγάλου όγκου και ετερογενών συλλογών

# Από την κωδικοποίηση στα εννοιολογικά σχήματα

- Οι τεχνολογικές εξελίξεις και η ανάγκη επικοινωνίας στο ψηφιακό περιβάλλον δημιούργησε νέες προκλήσεις για παραγωγή μεταδεδομένων και εντοπισμό/ταυτοποίηση πόρων, βασισμένων στις ανάγκες αναζήτησης των χρηστών
  - Έμφαση στην πληροφορία:
    - Τι πληροφορία στοχεύει να παρέχει μια βιβλιογραφική εγγραφή;
    - Ποιες πληροφοριακές ανάγκες των χρηστών θα μπορεί να καλύπτει;
- Ανάγκη ανάπτυξης μοντέλων δεδομένων
  - FRBR: Μοντέλο οντοτήτων - συσχετίσεων
    - Υπάρχουν διακριτές οντότητες (το μοντέλο ορίζει δέκα)
    - Κάθε οντότητα έχει μια σειρά από γνωρίσματα (attributes)
    - Οι οντότητες συσχετίζονται μεταξύ τους
  - Νέοι κανόνες καταλογογράφησης (RDA)
  - Νέα θεώρηση του βιβλιογραφικού καταλόγου
    - Δεν είναι μια σειρά από εγγραφές, αλλά ένα δίκτυο από συνδεμένα δεδομένα που καθιστούν ικανό το χρήστη να εκτελεί ομοιόμορφα όλες τις αναγκαίες λειτουργίες

# Από τα εννοιολογικά σχήματα στον παγκόσμιο ιστό

- Οι βιβλιοθήκες και ο «έξω» κόσμος: ανάγκη για διεύρυνση της διαλειτουργικότητας
  - Μεταδεδομένα από ειδικούς, από αυτόματες διαδικασίες, από χρήστες οι οποίοι μπορούν πλέον να καταθέτουν και σχολιασμούς
  - Οι βιβλιογραφικές εγγραφές μπορούν να διαλειτουργούν και να (επανα)χρησιμοποιούνται από άλλες εφαρμογές που λειτουργούν στο περιβάλλον του παγκόσμιου ιστού;
- Το ευρύτερο περιβάλλον:
  - Ολοκλήρωση δεδομένων - ανάκτηση πληροφορίας από διαφορετικές πηγές που φιλοξενούν ετερογενείς πληροφορίες.
  - Η εξελίξεις στο σημασιολογικό ιστό: εννοιολογική διαχείριση και διασύνδεση των πληροφοριών που διατίθενται μέσω του παγκόσμιου ιστού.

# Η νέα πρόκληση

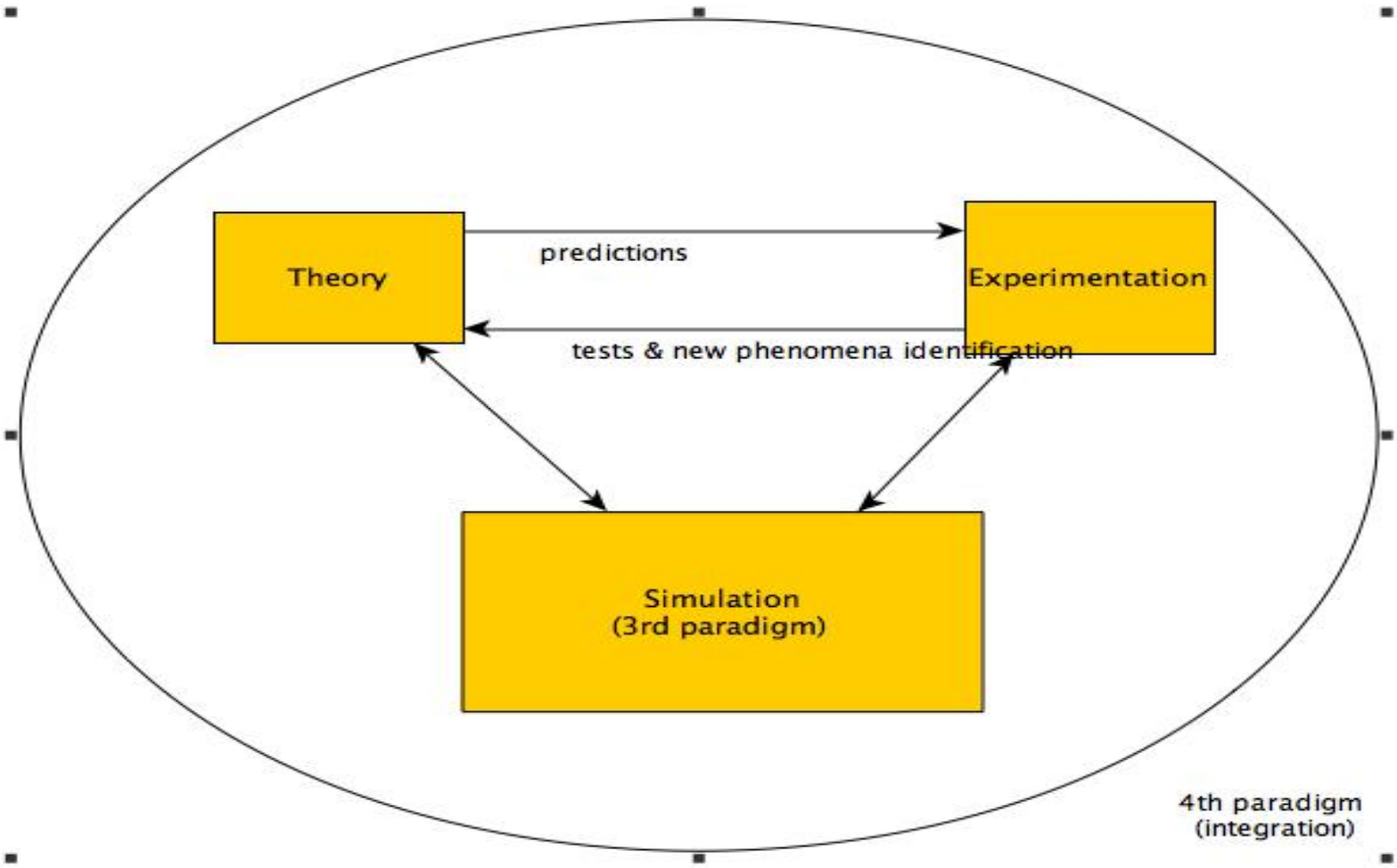
- Από τον ιστό των τεκμηρίων στον ιστό των δεδομένων
- Linked Open Data: Δεδομένα σε μορφή που διευκολύνεται η ανάπτυξη συνδέσμων μεταξύ πόρων συλλογών, στοιχείων μεταδεδομένων και όρων λεξιλογίων.
- Να δημιουργήσουμε συνδέσεις δηλωμένου τύπου (typed links) μεταξύ των δεδομένων από διαφορετικές πηγές του ιστού
- Πλεονεκτήματα
  - Διαμοιραζόμενα, επεκτάσιμα και επαναχρησιμοποιήσιμα δεδομένα προς όλες τις υπηρεσίες διαχείρισης δεδομένων και χρηστών με υποστήριξη πολυγλωσσίας
  - Οι πόροι μπορούν να περιγράφονται συνεργατικά, δυνατότητα πολλών περιγραφών στο ίδιο περιεχόμενο/αντικείμενο
  - Σύνδεση με δεδομένα άλλων κοινοτήτων ή προσώπων και να επαυξάνονται σύμφωνα με την εμπειρία του κάθε χρήστη
  - Δυνατότητες επιμέλειας: διεύρυνση συλλογής, επαύξηση της αξίας των πόρων, διεύρυνση προσβασιμότητας και αναφοράς των δεδομένων
  - Ανεξαρτησία από Integrated Library System

# Ιδιότητες του Ιστού των Δεδομένων

- Ο ιστός των Δεδομένων είναι ένα παραπάνω επίπεδο του κλασικού Ιστού των Τεκμηρίων, κάνοντας κοινή χρήση αρκετών ιδιοτήτων, όπως:
  - Ο Ιστός των Δεδομένων είναι γενικός και είναι δυνατόν να περιλαμβάνει οποιοδήποτε τύπο δεδομένων
  - Οποιοσδήποτε μπορεί να δημοσιεύσει δεδομένα στον Ιστό των Δεδομένων
  - Οι εκδότες δεδομένων δεν έχουν κανένα περιορισμό για τα λεξιλόγια που θα χρησιμοποιήσουν στην αναπαράσταση των δεδομένων τους
  - Οι οντότητες συνδέονται με RDF συνδέσμους, δημιουργώντας ένα καθολικό γράφο δεδομένων ο οποίος εκτείνεται σε πηγές δεδομένων και ενεργοποιεί την ανακάλυψη νέων πηγών δεδομένων

# Ερευνητικά δεδομένα – ακαδημαϊκή επικοινωνία

Εποχή των ερευνητικών δεδομένων - data intensive science



# Βασικές απαιτήσεις - χαρακτηριστικά

- Συστατικά: Βιβλιογραφία, Δεδομένα και Λογισμικό
- Ανταλλαγή υποθέσεων, αποτελεσμάτων, σκέψεων μεταξύ των ερευνητών ανεξαρτήτως τόπου και χρόνου
- Καθιέρωση κοινής ορολογίας, κοινοτήτων – προσώπων για την ανάπτυξη ερευνητικών πεδίων και τη συσχέτιση συναφών εργασιών
- Μεγάλης κλίμακας συνεργασία ανεξαρτήτως τόπου και χρόνου
- Μέσα για διαχείριση, τεκμηρίωση και επίλυση ασυμφωνιών
- Καθιέρωση προτεραιοτήτων
- Απαίτηση για βεβαιότητα της ποιότητας δεδομένων και αποτελεσμάτων (μέσω ετεροαναφορών και βιβλιομετρικών δεικτών)

# Προκλήσεις

- Βιβλιογραφία: τεράστια
- Αρχικά δεδομένα + αποτελέσματα προσομοίωσης: τεράστια
- Λογισμικό: τεράστιοι όγκοι εξαιτίας της έκρηξης των προσομοιωμένων πειραμάτων
- Αναφορές στα δεδομένα (data citations) από ανθρώπους ή μηχανές:
  - Τεράστιο πλήθος
  - Προκαλούν νέους υπολογισμούς

# Κατευθύνσεις

- Εργαλεία υποστήριξης ερευνητή:
  - Πρότυπα αναφοράς δεδομένων και συλλογών δεδομένων (data citation), πλοήγηση σε βιβλιογραφία ή από βιβλιογραφία σε δεδομένα
  - Περιβάλλοντα συνεργασίας, συγγραφής, σχολιασμού βιβλιογραφίας, προσομοίωσης και ανάλυσης
- Διαχείριση κειμένων και εξόρυξη γνώσης από αυτά
- Διαχείριση συνδεδεμένων δεδομένων
- Εξατομικευμένες υπηρεσίες πληροφόρησης και έρευνας (π.χ. συστάσεις για επιστημονικές υποθέσεις)

# Φάσεις διαχείρισης ερευνητικών δεδομένων

- Παρόμοιες με τη διαχείριση των δημοσιευμάτων
- Μητρώα δεδομένων
- Πιστοποίηση εγκυρότητας δεδομένων
- Ενημέρωση κοινού
- Αρχειοθέτηση και Διατήρηση
- Αναφορές στα δεδομένα

# Στόχος: να κάνουμε χρήσιμα τα δεδομένα

- Δομή: καλά ορισμένα μορφότυπα και σχήματα δεδομένων, καθολικοί κωδικοί ταυτοποίησης και λεξιλόγια/οντολογίες
- Προσιτά στους χρήστες: free-text tags, HTML microformats
- Αυτόματοι μηχανισμοί Εξόρυξης Γνώσης
- Ολοκλήρωση δομημένων δεδομένων με ημιδομημένα κείμενα
- Δομημένη πληροφορία από τις δημοσιεύσεις - nanopublications
- Επιμέλεια βάσεων δεδομένων
- Σύνδεση μεταξύ δεδομένων και δημοσιεύσεων, σε τέτοιο βαθμό που να μην έχει νόημα η διάκριση μεταξύ δεδομένων και δημοσιευμάτων

# Διαλειτουργικότητα

- Σημασιολογική διαλειτουργικότητα μεταξύ τεκμηρίων και δεδομένων με σκοπό την ενιαία πρόσβαση σε διαφορετικές πηγές δεδομένων και τεκμηρίων
- Συνδεδεμένα δεδομένα (linked data)
- Εξόρυξη εννοιών και γνώσης από κείμενα (text mining)
- Συνδυασμός και των δυο: εντοπισμός προσώπων, οργανισμών, τοποθεσιών, αναφορών (citations) και ερευνητικών δεδομένων σε σώματα δημοσιεύσεων

# Παράδειγμα 1: Στατιστική Ανάλυση σε Συνδεδεμένα Δεδομένα (1/2)

- Βήματα στατιστικής επεξεργασίας:
  - Συγκέντρωση στατιστικών δεδομένων από διάφορες πηγές (κυβέρνηση, οργανισμοί, αρχεία, βιβλιοθήκες)
  - Καθαρισμός και ομογενοποίηση δεδομένων για να είναι συγκρίσιμα
  - Πολύμορφη στατιστική επεξεργασία, ανάλυση και εξαγωγή συμπερασμάτων
- Πολλοί δημόσιοι οργανισμοί παρέχουν στατιστικά στοιχεία υπό μορφή linked data (Eurostat)
  - Τα linked data είναι εμπλουτισμένα δεδομένα λόγω της διασύνδεσής τους

# Παράδειγμα 1: Στατιστική Ανάλυση σε Συνδεδεμένα Δεδομένα (2/2)

- Ζητούμενο: Συσσώρευση και ολοκλήρωση δεδομένων από διαφορετικές πηγές linked data και ταυτόχρονη στατιστική επεξεργασία με επερωτήσεις SPARQL
- Πρόβλημα:
  - Λεξιλόγια για την περιγραφή πηγών συνδεδεμένων δεδομένων
  - Μονάδες μέτρησης
- Zapilko, B., & Mathiak, B. (2011). Performing Statistical Methods on Linked Data. International Conference on Dublin Core and Metadata Applications, 116-125. http://dcpapers.dublincore.org/pubs/article/view/3627/1853

# Παράδειγμα 2: Ταίριασμα Συνόλων Συνδεδεμένων Δεδομένων (1/2)

- Προσδιορισμός σημασιολογικών αντιστοιχιών μεταξύ σχημάτων ή μοντέλων μεταδεδομένων, βάσεων δεδομένων και εννοιολογικών μοντέλων (π.χ. οντολογιών)
- Συνεντεύξεις για τον τρόπο «κατανάλωσης» συνδεδεμένων δεδομένων στις κοινωνικές επιστήμες
- Καλές πρακτικές για τη δημοσίευση συνδεδεμένων δεδομένων στις κοινωνικές επιστήμες
- Δημιουργία/Τροποποίηση οντολογιών για την αναπαράσταση κοινοτήτων, δεδομένων και λεξιλογίων

# Παράδειγμα 2: Ταίριασμα Συνόλων Συνδεδεμένων Δεδομένων (2/2)

- Κριτήρια για από κοινού ανάλυση και επεξεργασία δύο συνόλων δεδομένων σε περιβάλλον ιστού
  - Σε επίπεδο αντικειμένου (instance – based schema matching)
  - Σε επίπεδο ιδιοτήτων των αντίστοιχων οντολογιών (object property matching utilizing the overlap between Imported ontologies)
- Zapiko, B. (2015) Methods for Matching of Linked Open Social Science Data, PhD thesis, https://ub-madoc.bib.uni-mannheim.de/37467/1/Dissertation_Zapilko.pdf

# Προβληματισμός

Θα συνδέσουμε τα δεδομένα των βιβλιοθηκών με τα ερευνητικά δεδομένα;

# Η απάντηση στη νέα πρόκληση

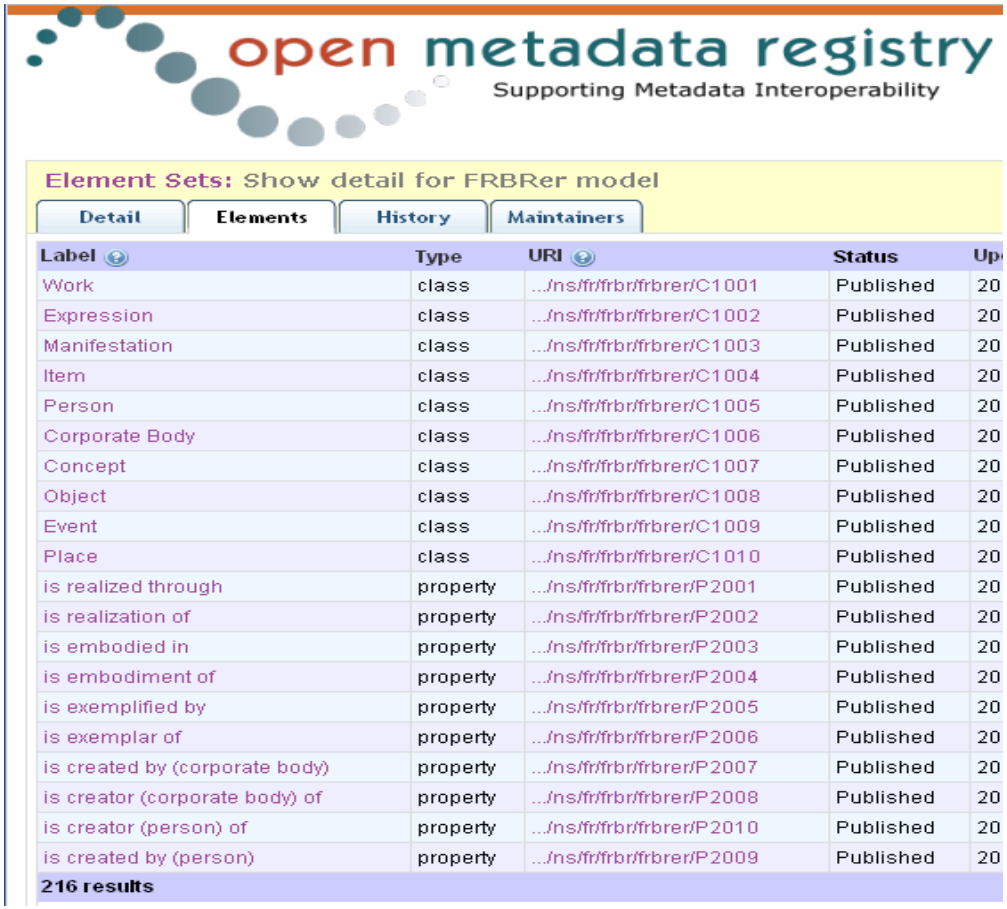
- Βιβλιοθήκη
  - μια συλλογή από φυσικά ή αφηρημένα αντικείμενα (συμπεριλαμβανομένων των ψηφιακών
  - Μια τοποθεσία που βρίσκεται η συλλογή
  - Ένας φορέας (πρόσωπο ή οργανισμός) που επιμελείται τη συλλογή και διαχειρίζεται την τοποθεσία
  - Όλο το φάσμα των οργανισμών πολιτιστικής κληρονομιάς
- Δεδομένα βιβλιοθήκης
  - Κάθε τύπος ψηφιακής πληροφορίας που παράγεται ή επιμελείται από βιβλιοθήκες και περιγράφει πόρους ή βοηθά στην ανακάλυψή τους. Τα δεδομένα βιβλιοθήκης αφορούν σε πόρους συλλογών, στοιχεία μεταδεδομένων και όρων λεξιλογίων
- Προς ένα νέο βιβλιογραφικό πλαίσιο
  - Αξιοποίηση μοντέλων δεδομένων και κανόνων διαχείρισης περιεχομένου, συμπεριλαμβανομένου του MARC
  - Διαλειτουργικότητα λεξιλογίων και αρχείων καθιερωμένων όρων
  - Συνδέσεις μεταξύ καταλόγων, με βάση τα θέματα, ονόματα, τοπωνύμια κ.α
  - Αναφορές σε βιβλιογραφικά δεδομένα (citations) και κοινωνικές επισημειώσεις, σχολιασμοί κλπ
  - Ανακάλυψη συλλογών χρήσιμων για μια κοινότητα
  - Πλοήγηση, αναζήτηση, ανάκτηση και παρουσίαση πληροφορίας στο νέο περιβάλλον

# Συνδεδεμένα Δεδομένα Βιβλιοθηκών (Library Linked Data)

- Κύρια αποτελέσματα
  - Έκθεση
    - http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/
  - Έκθεση με εφαρμογή τεχνολογιών Σημασιολογικού Ιστού στις Βιβλιοθήκες και σε συναφής τομείς
    - http://www.w3.org/2005/Incubator/lld/XGR-lld-usecase-20111025/
  - Έκθεση για Σύνολα δεδομένων (Datasets), Τιμές λεξιλογίων (Value vocabularies), Σύνολα Στοιχείων Μεταδεδομένων (Metadata element sets or element sets)
    - http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

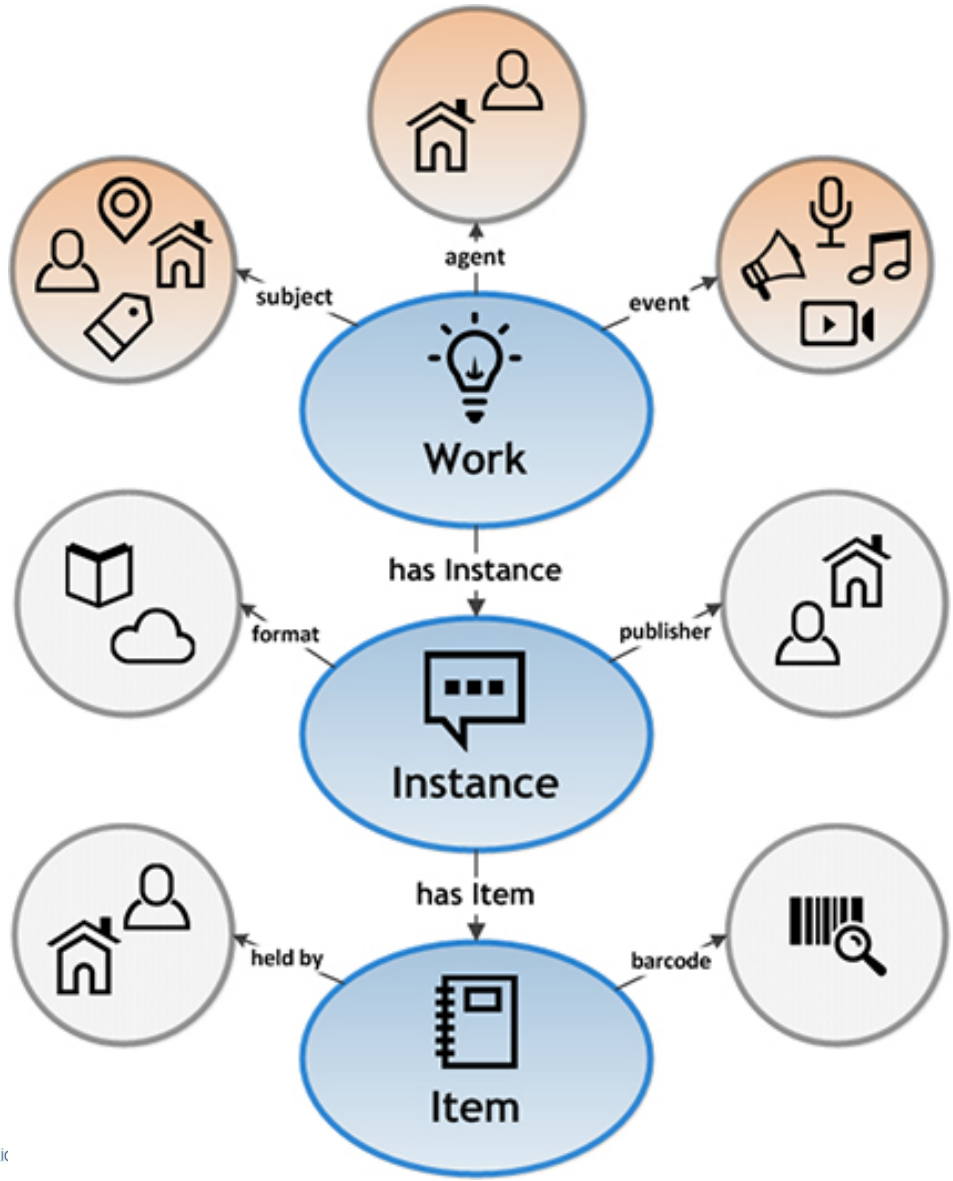
# Λεξιλόγια – Βιβλιογραφικά δεδομένα

- The FRBR element set vocabulary
  - http://iflastandards.info/ns/fr/




| Label | Type | URI | Status | Up |
|---|---|---|---|---|
| Work | class | .../ns/fr/frbr/frbrer/C1001 | Published | 20 |
| Expression | class | .../ns/fr/frbr/frbrer/C1002 | Published | 20 |
| Manifestation | class | .../ns/fr/frbr/frbrer/C1003 | Published | 20 |
| Item | class | .../ns/fr/frbr/frbrer/C1004 | Published | 20 |
| Person | class | .../ns/fr/frbr/frbrer/C1005 | Published | 20 |
| Corporate Body | class | .../ns/fr/frbr/frbrer/C1006 | Published | 20 |
| Concept | class | .../ns/fr/frbr/frbrer/C1007 | Published | 20 |
| Object | class | .../ns/fr/frbr/frbrer/C1008 | Published | 20 |
| Event | class | .../ns/fr/frbr/frbrer/C1009 | Published | 20 |
| Place | class | .../ns/fr/frbr/frbrer/C1010 | Published | 20 |
| is realized through | property | .../ns/fr/frbr/frbrer/P2001 | Published | 20 |
| is realization of | property | .../ns/fr/frbr/frbrer/P2002 | Published | 20 |
| is embodied in | property | .../ns/fr/frbr/frbrer/P2003 | Published | 20 |
| is embodiment of | property | .../ns/fr/frbr/frbrer/P2004 | Published | 20 |
| is exemplified by | property | .../ns/fr/frbr/frbrer/P2005 | Published | 20 |
| is exemplar of | property | .../ns/fr/frbr/frbrer/P2006 | Published | 20 |
| is created by (corporate body) | property | .../ns/fr/frbr/frbrer/P2007 | Published | 20 |
| is creator (corporate body) of | property | .../ns/fr/frbr/frbrer/P2008 | Published | 20 |
| is creator (person) of | property | .../ns/fr/frbr/frbrer/P2010 | Published | 20 |
| is created by (person) | property | .../ns/fr/frbr/frbrer/P2009 | Published | 20 |
| 216 results | | | | |

# BIBFRAME μοντέλο



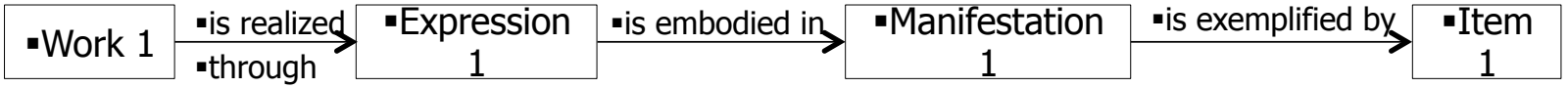
# Δεδομένα βιβλιοθηκών: Ενδεικτικές υλοποιήσεις σε ΣΔ

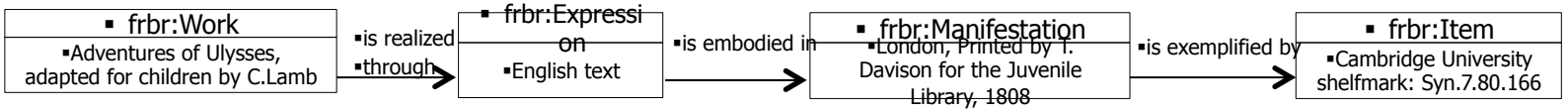
- Virtual International Authority File (VIAF)
  - Library of Congress Name Authority File (LC/NAF)
  - The BIBSYS personal name authority file
- Library of Congress Subject Headings (LCSH)

- The British National Bibliography (BnB)
  - http://www.bl.uk/bibliographic/datafree.html
- Η Biblioteca Nacional de España σε συνεργασία με το Ontology Engineering Group (OEG)
  - http://www.bne.es/en/Catalogos/DatosEnlazados/
  - Υλοποίηση συνδεδεμένων δεδομένων και FRBR εννοιολογικού μοντέλου
  - MARiMbA: Μεταξύ των αποτελεσμάτων του έργου ήταν και η ανάπτυξη ενός μετατροπέα από MARC 21 σε RDF

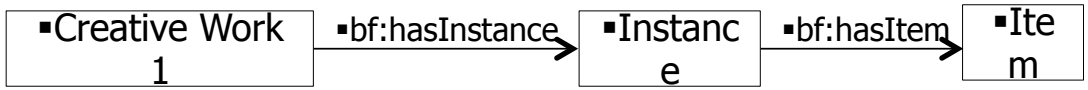
# Παραδείγματα (1/2)

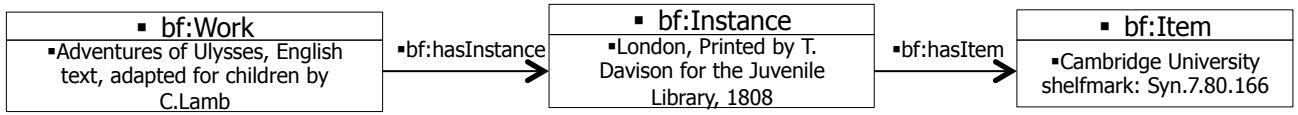
## FRBR



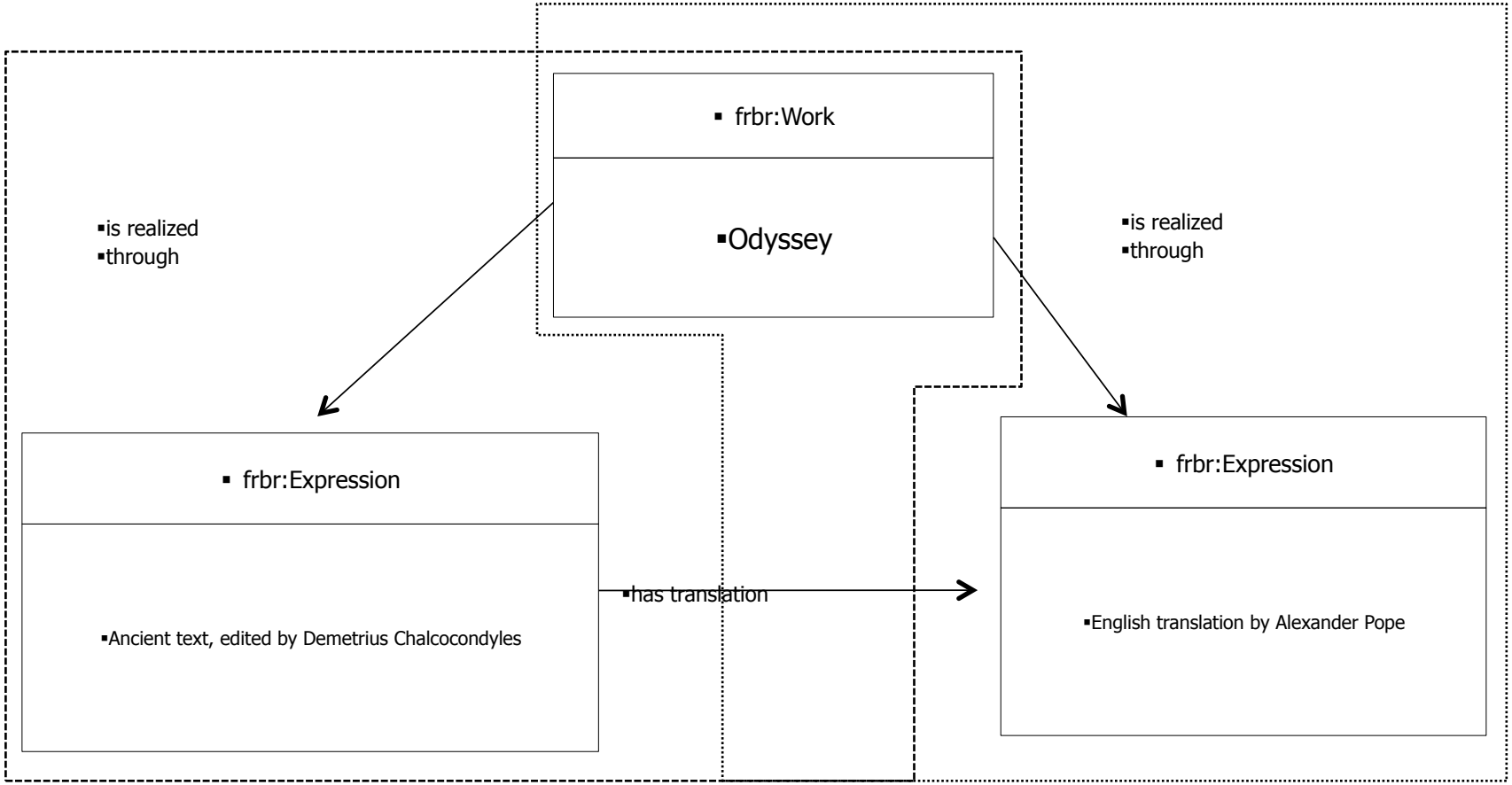

## BIBFRAME



# Παραδείγματα (2/2)

## FRBR Translation



# Virtual International Authority File (VIAF)

- Διεθνές Εικονικό Αρχείο Καθιερωμένων Τύπων για ονόματα προσώπων
- Συνδέσεις μεταξύ Αρχείων (Καθιερωμένων Τύπων)
  - Τα δεδομένα παραμένουν και τα συντηρούν οι φορείς που τα παράγουν
  - Σύνθεση με διαδικασίες συγκομιδής
- Πρόσβαση μέσα από τεχνολογίες του Ιστού
- Πολύγλωσσα Δεδομένα
- Πλήρης συμβατότητα με τα υπάρχοντα πρότυπα (MARC21, UNIMARC)
- Επεκτάσιμο για οποιοδήποτε αριθμό εθνικών αρχείων καθιερωμένων όρων

# VIAF – υλοποίηση

- Πρόβλημα ταύτισης
  - Διαφορετικά Πρόσωπα – Ένα Όνομα
  - Ένα Πρόσωπο – Πολλά Ονόματα

- Απόφαση για δημιουργία Αναβαθμισμένων Καθιερωμένων Τύπων

# VIAF – ταύτιση

- Η ταύτιση γίνεται με τους Αναβαθμισμένους Καθιερωμένους Τύπους