Μοντελοποίηση Δεδομένων

XML
Προδιαγραφές Τεκμηρίωσης
Δομημένα Λεξιλόγια

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

Επικοινωνία: janag@dib.uth.gr

www.anagnostopoulos.name

XML

- Extensible markup language – Μία απλή έκδοση της sgml
- Σχεδιασμένη για να εισάγει την sgml στο διαδίκτυο
- Μια μετα-γλώσσα (meta-language)
  - Μια γλώσσα που παράγει γλώσσες

Αναμενόμενα κέρδη

- Αποθήκευση μια φορά και μορφοποίηση πολλές φορές
- Ανεξαρτησία υλικού – λογισμικού
- Συγκέντρωση δεδομένων μια φορά και ανταλλαγή πολλές φορές
- Ταχύτερη εστιασμένη αναζήτηση
- Μικρότερη συμφόρηση του δικτύου
- Οι μηχανές αναζήτησης αναζητούν συγκεκριμένες ετικέτες (tags) στον κώδικα XML
  - Ταχύτερα
  - Με μεγαλύτερη ακρίβεια

HTML

- Χρησιμοποιεί tags ανάμεσα στο κείμενο...
  - ...για να περιγράψει το layout της σελίδας

```
<p> <b>Alan</b>, 42 years, <i>agb@abc.com</i>
```

- Σχεδιάστηκε ειδικά για να περιγράψει την παρουσίαση και όχι το περιεχόμενο

XML

- Σχεδιάστηκε ειδικά για να περιγράψει το περιεχόμενο (content) και όχι την παρουσίαση μιας σελίδας
- Βασικές διαφορές από την HTML
  - Μπορεί κανείς να ορίσει νέα tags κατά βούληση
  - Nested tags σε οποιοδήποτε βάθος
  - Ενα έγγραφο XML μπορεί προαιρετικά να περιέχει μια περιγραφή της γραμματικής του
- Τα tags δομούν το περιεχόμενο

```
<person><name>Kostas</name>…</person>
```

  - Το πως θα εμφανιστούν ορίζεται ξεχωριστά από κάποιο stylesheet (XSL)

XML

- **Ο ρόλος της XML**
  - **Προτάθηκε σαν μια** markup γλώσσα **περιγραφής εγγράφων**
  - **Καταγωγή απο την** SGML **(ψηφιακές βιβλιοθήκες)**
  - **Εξελίσσεται όμως σε ένα παγκόσμιο πρότυπο για** ανταλλαγή πληροφορίας

- **Βασικό συστατικό της XML είναι το** element
  - **Τιμή του element → Κείμενο που περικλείεται από ένα ζεύγος tags**
  - **Εκτός από την** λογική **δομή (elements), τα tags περιγράφουν και την** φυσική **δομή** (entities)

```
<person>
   <name>Alan</name>
   <age>42</age>
   <email>alan@abc.com</email>
   <email>abrown@mail.com</email>
</person>
```

example

Book Catalogue

id

author

title

price

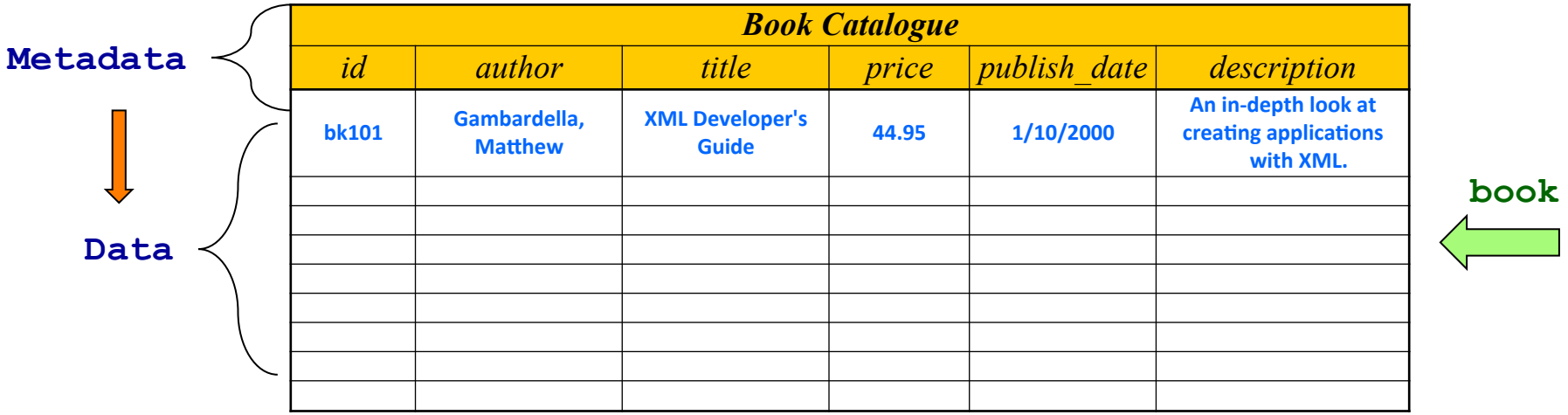
publish_date

description

Τι κάνουμε εδώ;

Excel, DB, ή κάτι… άλλο;

Πλεονεκτήματα / Μειονεκτήματα;




```
<Book Catalogue>
    <book>
        <id>bk101</id>
        <author>Gambardella, Matthew</author>
        <title>XML Developer's Guide</title>
        <price>44.95</price>
        <publish_date>01/10/2000</publish_date>
        <description>An in-depth look at creating applications with XML.</description>
    </book>
</Book Catalogue>
```

Book Catalogue					
id	*author*	*title*	*price*	*publish_date*	*description*
bk101	Gambardella, Matthew	XML Developer's Guide	44.95	01/10/2000	An in-depth look at creating applications with XML.
817525766-0	Αναγνωστόπουλος, Ιωάννης	Τεχνολογίες Διαδικτύου	88.00	01/03/2009	Βασικές αρχές τεχνολογιών Διαδικτύου και προγραμματισμού.

<Book Catalogue>

<book>

<id>**bk101**</id>

<author>**Gambardella, Matthew**</author>

<title>**XML Developer's Guide**</title>

<price>**44.95**</price>

<publish_date>**01/10/2000**</publish_date>

<description>**An in-depth look at creating applications with XML.** </description>

</book>

<book>

<id>**817525766-0**</id>

<author>**Αναγνωστόπουλος, Ιωάννης**</author>

<title>**Τεχνολογίες Διαδικτύου**</title>

<price>**88.00**</price>

<publish_date>**01/03/2009**</publish_date>

<description>**Βασικές αρχές τεχνολογιών Διαδικτύου και προγραμματισμού.**</description>

</book>

</Book Catalogue>

Book Catalogue					
book_id (format)	*author (language)*	*title (language)*	*price (currency)*	*publish_date (calendar type)*	*description (language)*
bk101	Gambardella, Matthew	XML Developer's Guide	44.95	01/10/2000	An in-depth look at creating applications with XML.
817525766-0	Αναγνωστόπουλος, Ιωάννης	Τεχνολογίες Διαδικτύου	88.00	01/03/2009	Βασικές αρχές τεχνολογιών Διαδικτύου και προγραμματισμού.

(More Metadata inside the parentheses)
οπότε πως είναι ο κώδικάς μου…;

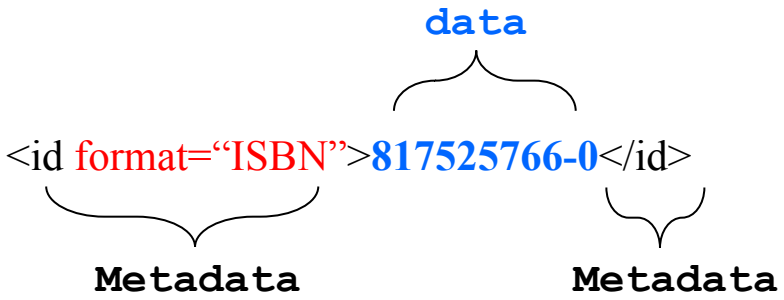
Πρέπει να περιγράψουμε τα data που περιγράφουν
τα metadata
(meta-metadata;)

Π.χ.το book_id format ενός βιβλίου είναι το ISBN

Για να δούμε πως γράφεται σε XML αυτό…

<id format="ISBN">**817525766-0**</id>

Book Catalogue					
id (format)	*author (language)*	*title (language)*	*price (currency)*	*publish_date (calendar type)*	*description (language)*
bk101	Gambardella, Matthew	XML Developer's Guide	44.95	01/10/2000	An in-depth look at creating applications with XML.
817525766-0	Αναγνωστόπουλος, Ιωάννης	Τεχνολογίες Διαδικτύου	88.00	01/03/2009	Βασικές αρχές τεχνολογιών Διαδικτύου και προγραμματισμού.



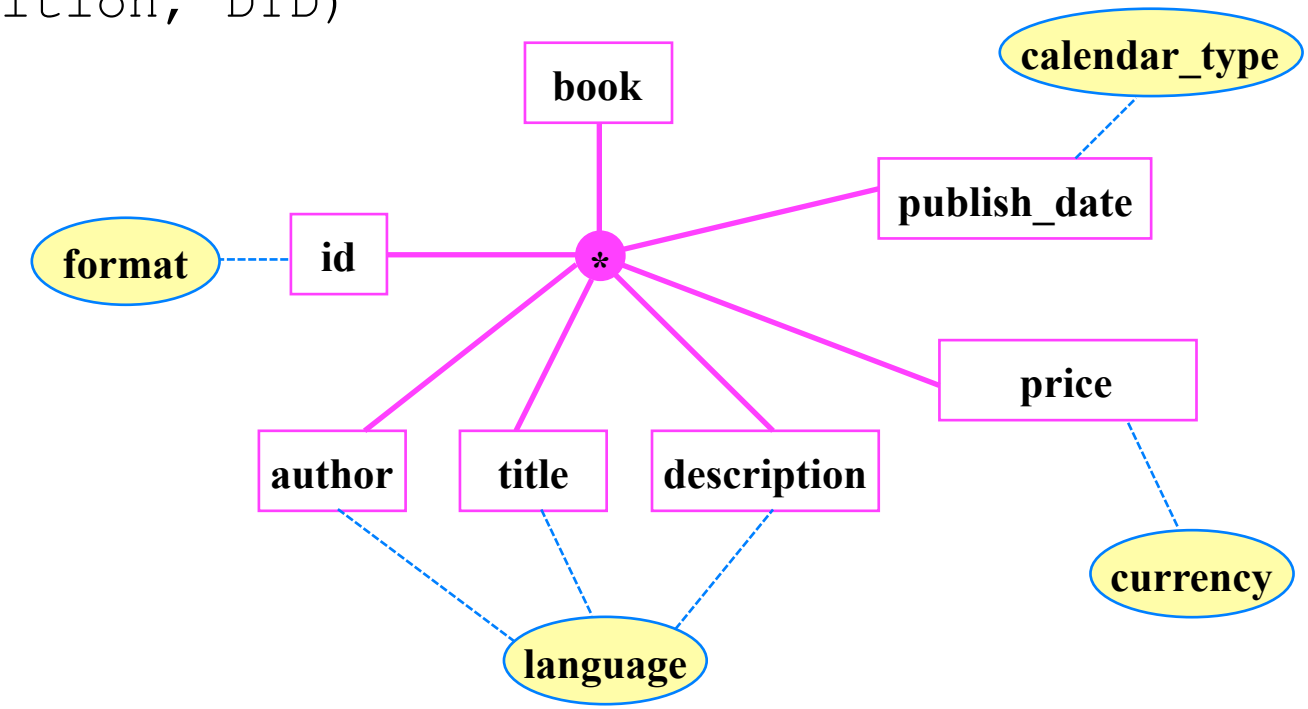
id: element
format:attribute_of_element(id)
ISBN: value_of_attribute(format)

817525766-0: value_of_element(id)

```
<Book Catalogue>
    <book>
        <id>bk101</id >
        <author language="English" >Gambardella, Matthew</author>
        <title language="English">XML Developer's Guide</title>
        <price currency="USD" >44.95</price>
        <publish_date calendar_type="Gregorian Little Endian" >01/10/2000</publish_date>
        <description language="English">An in-depth look at creating applications with XML. </description>
    </book>
    <book>
        <id format="ISBN">817525766-0</id>
        <author language="Greek" >Αναγνωστόπουλος, Ιωάννης</author>
        <title language="Greek" >Τεχνολογίες Διαδικτύου</title>
        <price currency="Euro">88.00</price>
        <publish_date calendar_type="Gregorian Little Endian" >01/03/2009</publish_date>
        <description language="Greek" >Βασικές αρχές τεχνολογιών Διαδικτύου και προγραμματισμού.</description>
    </book>
</Book Catalogue>
```

Schema

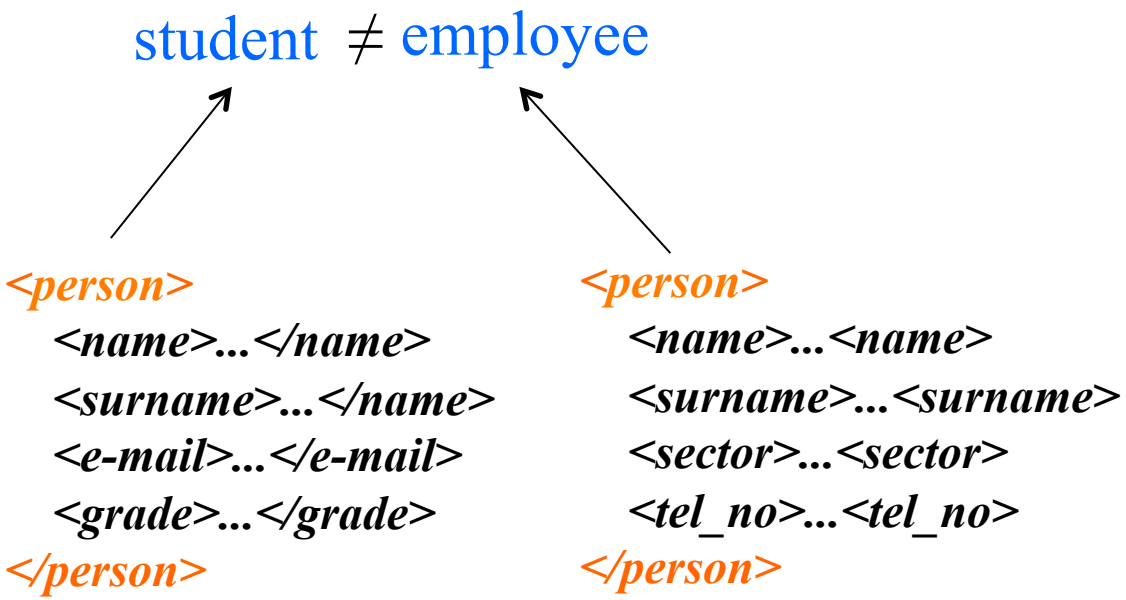
(Document Type Definition, DTD)



XML Namespaces

- Τα ονόματα των elements ενός αρχείου χαρακτηρίζονται από μια ετικέτα (super label)
- Το label αυτό ονομάζεται *namespace* και αποτελεί το όνομα της συλλογής των ονομάτων του αρχείου
- Το όνομα ενός namespace παίρνει τη μορφή ενός URI, το δεν έχει link πουθενά! Χρησιμοποιείται αυτή η αναπαράσταση γιατί είναι μοναδική
- *συνήθως μοιάζει με ένα URL (διεύθυνση Διαδικτυακού Τόπου)*

Παράδειγμα name conflict



Name Conflicts

Εφόσον τα ονόματα των element στην XML δεν είναι προκαθορισμένα (δεσμευμένα), ενδέχεται δύο διαφορετικά XML documents να χρησιμοποιούν το ίδιο element name

Λύση … τα XML Namespaces

Το XML namespace attribute τοποθετείται στο αρχικό tag ενός element και έχει την παρακάτω σύνταξη:

xmlns:namespace-prefix="namespaceURI" ή

xmlns=namespaceURI“ (obsolete)

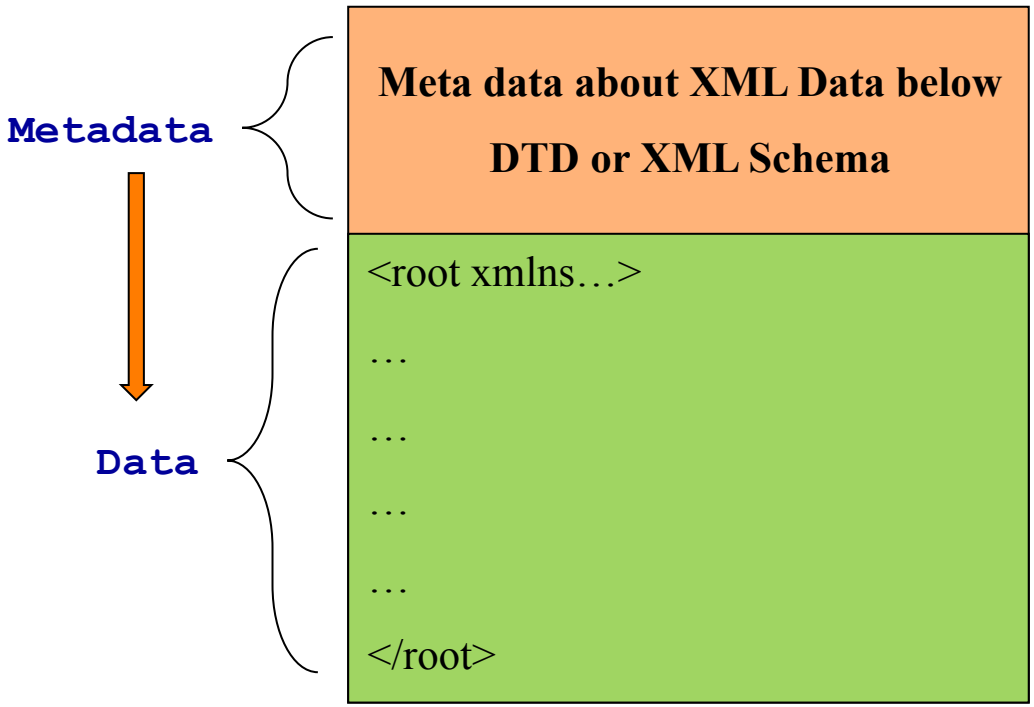
Όταν ένα namespace καθορίζεται στο αρχικό tag ενός element, όλα τα child elements με το ίδιο namespace-prefix συνδέονται με το ίδιο namespace

Λύση του name conflict με χρήση Namespace

```
<person xmlns:janag="http://www.anagnostopoulos.name/my_xmlns#" >
   <janag:name>...</janag:name>
   <janag:surname>...</janag:surname>
   <janag:e-mail>...</janag:e-mail>
   <janag:grade>...</janag:grade>
</person>
```

```
<person xmlns:p="http://gr.linkedin.com/pub/agissilaos-papantoniou/4/209/758/agis_xmlns#">
   <p:name>...</p:name>
   <p:surname>...</p:surname>
   <p:sector>...</p:sector>
   <p:tel_no>...</p:tel_no>
</person>
```

Δομή αρχείου XML (instance)



XML in Digital Culture – Ένα απλό παράδειγμα

```
<exhibit>
   <name language="Italian">La Gioconda</name>
   <creator nationality="Italian">Leonardo da Vinci</creator>
   <museum language="French">
      <name>Musée du Louvre</name>
      <country>France</country>
      <city>Paris</city>
      <zip>75058</zip>
   </museum>
</exhibit>
```

Ποιο το σχήμα εδώ;

Προδιαγραφές Τεκμηρίωσης

Μια ενδεικτική προσέγγιση

Διατήρηση ψηφιακών πόρων

- Η διατήρηση πληροφορίας σε ψηφιακή μορφή αποτελεί πρόβλημα αυξανόμενης σημασίας, που αντιμετωπίζουν ιδίως οργανισμοί στο τομέα της πολιτιστικής κληρονομιάς
  - Μουσεία, ψηφιακές βιβλιοθήκες, ψηφιακά αρχεία
- Η τήρηση κατάλληλων μεταδεδομένων αποτελεί το 'κλειδί' για την διατήρηση ψηφιακών αντικειμένων

Διαδικασίες ψηφιακής διατήρησης

- Ορίζεται ως οποιαδήποτε ενέργεια διατήρησης που εκτελείται πάνω σε ένα ψηφιακό αντικείμενο
- Παραδείγματα
  - Μετάβαση σε διαφορετικό μορφότυπο
  - Συμπίεση
  - Έλεγχος ακεραιότητας ψηφιακού περιεχομένου
- Τα μεταδεδομένα διατήρησης χρησιμοποιούνται για την καταγραφή των σχετικών λεπτομερειών, π.χ:
  - Τύπος της διαδικασίας
  - Αποτελέσματα
  - Αναφορά σε εμπλεκόμενα πρόσωπα
    - Αναφορά στην σχετική άδεια

Μεταδεδομένα διατήρησης

- Πληροφορία που υποστηρίζει και τεκμηριώνει διαδικασίες που σχετίζονται με την ψηφιακή διατήρηση
- Τα μεταδεδομένα διατήρησης παρέχουν την απαραίτητη πληροφορία που εξασφαλίζει βιωσιμότητα, δυνατότητα παρουσίασης και δυνατότητα ερμηνείας των ψηφιακών πόρων
- Παραδείγματα
  - Αναγνωριστικό αντικειμένου
  - Θέση αποθήκευσης
  - Λογισμικό και υλικό που απαιτείται για την πρόσβαση στο αντικείμενο
  - Μορφότυπος του αντικειμένου
  - Τεχνικές λεπτομέρειες

Μεταδεδομένα διατήρησης: Κεντρικές οντότητες

1. Αντικείμενα
2. Γεγονότα
3. Διάφοροι Παράγοντες (περιβάλλον χώρος, περιορισμοί)
4. Δικαιώματα (πρόσβασης ή/και πνευματικά)

Αντικείμενα

- **Αναγνωριστικό αντικειμένου** (Τιμή, Μέθοδος, Τύπος)
- **Θέση αποθήκευσης**
- **Όνομα πρωτοτύπου**
- **Επίπεδο διατήρησης** (αρχείο, ροή δυαδικών ψηφιών, συνδυασμός κ.α.)
- **Φυσικά χαρακτηριστικά**
  - **Μορφότυπος** (Όνομα, Έκδοση)
  - **Μέγεθος**
  - **Παγιότητα** (Μέθοδος ελέγχου, Τιμή ελέγχου, Αρχική τιμή)
  - **Απαγορεύσεις**
  - **Σημαντικές ιδιότητες**
  - **Εφαρμογή δημιουργίας ψηφιακού πόρου**
  - **Περιβάλλον απαραίτητο για την πρόσβαση στο ψηφιακό πόρο**
    - Υλικό (Απαιτήσεις, Τεκμηρίωση)
    - Λογισμικό (Τύπος, Όνομα, Έκδοση, κ.λπ.)

?

Αντικείμενο (συνέχεια)

- **Τεχνικά μεταδεδομένα**
  - **Κείμενο** (Σύνολο χαρακτήρων/λέξεων/σελίδων)
  - **Εικόνα** (Ανάλυση, Διαστάσεις, Συμπίεση, Χρωματικό μοντέλο)
  - **Αρχείο Ήχου** (Ανάλυση, Διάρκεια, Συμπίεση, Κανάλια)
  - **Αρχείου Βίντεο** (Διαστάσεις πλαισίου, Διάρκεια, Αριθμός πλαισίων, Μέθοδος Codec)

Γεγονότα

Πληροφορία που καταγράφει την ιστορία/εξέλιξη (versioning) ενός ψηφιακού αντικειμένου

Κάθε γεγονός περιγράφει μια διαδικασία που τα αποτελέσματα της έχουν κάποια επιρροή

- **Μεταδεδομένα**
  - **Αναγνωριστικό** (Τιμή, Σχήμα)
  - **Τύπος γεγονότος** (π.χ. μετάβαση, συμπίεση, διαγραφή, έλεγχος εγκυρότητας)
  - **Αποτέλεσμα** (κωδικοποιημένες τιμές)
  - **Τεκμηρίωση αποτελέσματος**
  - **Τεκμηρίωση γεγονότος**
  - **Ημερομηνία/Ώρα**
  - **Συσχέτιση άδεια που επιτρέπει την εκτέλεση του γεγονότος**
  - **Συσχέτιση με παράγοντα (επόμενη διαφάνεια)**

Παράγοντες

Καταγραφή πληροφορίας σχετικά με παράγοντες (φυσικά πρόσωπα / φορείς) που σχετίζονται με γεγονότα διατήρησης και διαχείριση δικαιωμάτων

- **Ένας παράγοντας μπορεί να εκτελεί, να εξουσιοδοτεί ή να επιβάλλει ένα ή περισσότερα γεγονότα**
  - Δημιουργός / Φορέας Digital Culture
- **Ένας παράγοντας μπορεί να κρατάει ή να παραχωρεί ένα ή περισσότερα δικαιώματα**

- **Μεταδεδομένα**
  - **Αναγνωριστικό παράγοντα (π.χ. URI)**
  - **Όνομα παράγοντα**
  - **Ρόλος (Δημιουργός, Διαχειριστής, Νόμιμος Ιδιοκτήτης)**

Δικαιώματα

Πληροφορία που αναφέρεται στην άδεια που παρέχεται σε ένα σύστημα τεκμηρίωσης, ώστε να εκτελεστούν ενέργειες πάνω στα ψηφιακά αντικείμενα με σκοπό την διατήρηση

- **Παραδείγματα**
  - **Άδεια ανανέωσης ψηφιακού περιεχομένου**
  - **Άδεια μετατροπής μορφότυπου**
  - **Άδεια αναδημιουργίας του ψηφιακού περιεχομένου**

- **Μεταδεδομένα**
  - **Δήλωση άδειας**
  - **Συσχέτιση με παράγοντες**
  - **Συσχέτιση με αντικείμενα**
  - **Συμφωνητικό μεταβίβασης**

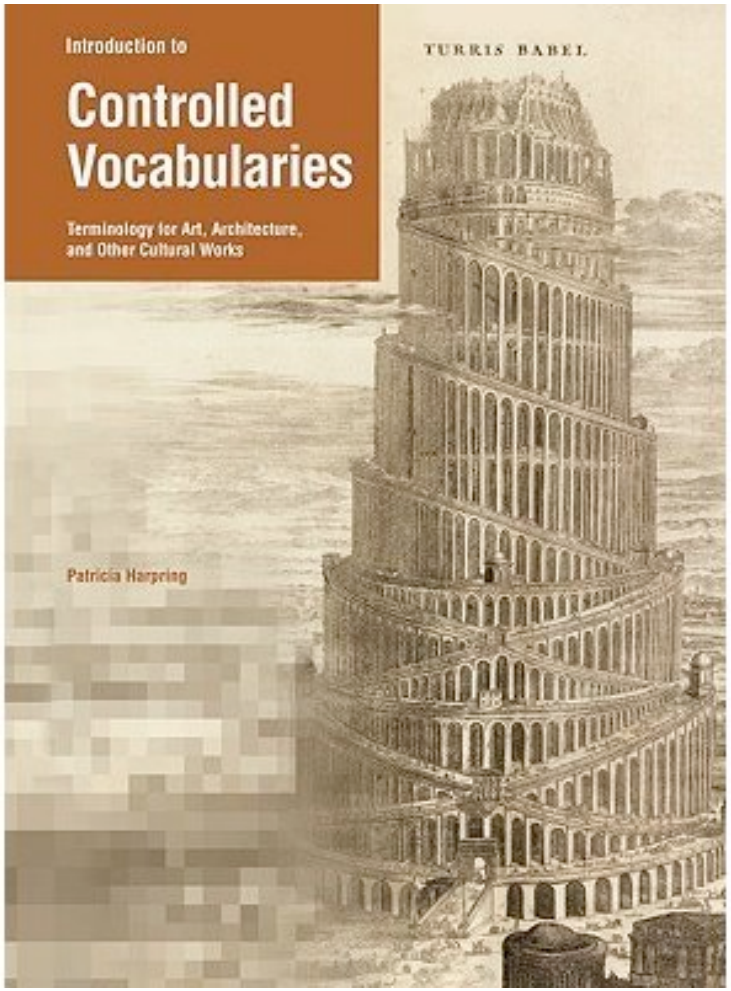
Διαδικασία συμπλήρωσης μεταδεδομένων

- Εξαγωγή μεταδεδομένων από το ίδιο το αντικείμενο,συνοδευόμενα έγγραφα
  - όνομα πρωτότυπου
  - μορφότυπος
  - περιβάλλον λογισμικού
  - μέγεθος
  - δικαιώματα
  - παράγοντες
- Εξαγωγή μεταδεδομένων κατά την διάρκεια δημιουργίας του ψηφιακού περιεχομένου
  - Τεχνικά μεταδεδομένα
- Αυτόματη συμπλήρωση από το σύστημα αρχειοθέτησης
  - Αναγνωριστικά (Αντικειμένου, Γεγονότος)
  - Θέση αποθήκευσης
  - Καταγραφή γεγονότος απόκτησης αντικειμένου

Δομημένα Λεξιλόγια

@wikipedia:

Controlled vocabularies provide a way to organize knowledge for subsequent retrieval. They are used in subject indexing schemes, subject headings, thesauri, taxonomies and other forms of knowledge organization systems. Controlled vocabulary schemes mandate the use of predefined, authorised terms that have been preselected by the designers of the schemes, in contrast to natural language vocabularies, which have no such restriction.



Δομημένα Λεξιλόγια: Παραδείγματα

- Library of Congress Subject Headings
  - https://en.wikipedia.org/wiki/Library_of_Congress_Subject_Headings
- Schema.org
  - https://schema.org/
- Dublin Core
  - http://dublincore.org/
  - https://en.wikipedia.org/wiki/Dublin_Core

References

- Αναγνωστόπουλος Ιωάννης, Τεχνολογίες Διαδικτύου, Πανεπιστήμιο Αιγαίου
- Γαβαλάς Δαμιανός, Τεχνολογίες Παγκόσμιου Ιστού, Πανεπιστήμιο Αιγαίου
- Αναγνωστόπουλος Ιωάννης, Προγραμματισμός στο Παγκόσμιο Ιστό, Πανεπιστήμιο Αιγαίου
- Μερόπη Πετράκη, Τεκμηρίωση και διατήρηση ψηφιακού περιεχομένου, ημερίδα «Πληροφοριακή Σχεδίαση για Πολιτισμική Τεκμηρίωση και Διαλειτουργικότητα»