arXiv's current and planned metadata formats

Simeon Warner - CS/CIS

simeon@cs.cornell.edu

Plan and aims

- A little history
- Present (why, what is good, shortcomings)
- Plans and motivations
- Ideas on new format (hoping for constructive comments/discussion)

In the beginning (1991)…

- Most of the metadata was not separated into fields -- the abstract and "the rest":

```
------------------------------------------------------------------------
\\
Paper: hep-th/9205071
From: dschwarz@email.tuwien.ac.at (Dominik SCHWARZ)
Date: Wed, 20 May 92 10:39:17 MET DST   (15kb)

Novel Symmetry of Non-Einsteinian Gravity in Two Dimensions, by H.Grosse, W.
Kummer, P. Pre\v{s}najder and D.J. Schwarz, 17 pages, TUW-92-04 (LaTeX)
Journal-ref: J.Math.Phys. 33 (1992) 3892-3900
\\
The integrability of $R^2$-gravity with torsion in two
dimensions is traced to an ultralocal dynamical symmetry of
constraints and momenta in Hamiltonian phase space. It may be

...
```

1995…now

- arXiv started collecting fielded metadata:

http://arXiv.org/ftp/hep-th/papers/9501/9501001.abs

- Mainly ASCII email dissemination (still important now)

```
------------------------------------------------------------------------
\\
Paper: hep-th/9501001
From: Fuad Saradzhev <fuad@yunus.mam.tubitak.gov.tr>
Date: Mon, 2 Jan 1995 14:42:25 +0200 (EET)   (24kb)

Title: Anomaly and Exotic Statistics in One Dimension
Author: Fuad Saradzhev
Comments: LATEX file, 38 pages
Report-no: TUBITAK preprint MRC--PH--TH.16--94, 1994.
\\
We study the influence of the anomaly on the physical quantum picture of the
chiral Schwinger model (CSM) defined on $S^1$. We show that such phenomena as
the total screening of charges and the dynamical mass generation
```

Internal and displayed

http://arXiv.org/ftp/hep-th/papers/9901/9901001.abs
http://arXiv.org/abs/hep-th/9901001

```
------------------
\\
Paper: hep-th/990100
From: "Yosuke Imamur
Date: Fri, 1 Jan 199
Date (revised v2): T
Date (revised v3): M

Title: String Juncti
Author: Yosuke Imamu
Comments: 13 pages +
Report-no: YITP-99-1
Journal-ref: Prog.Th
\\
  We explicitly give
theory compactified
compactified on $S^2
\\
```

[hep-th/9901001] String Junctions and Their Duals in Heterotic String Theory - Mozilla {Build ID: 2002052918}

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://arxiv.org/abs/hep-th/9901001 Go Search

Home Bookmarks WebMail Radio People Yellow Pages Download Calendar Channels

High Energy Physics - Theory, abstract hep-th/9901001

From: "Yosuke Imamura" <imamura@yukawa.kyoto-u.ac.jp>
Date (v1): Fri, 1 Jan 1999 01:01:10 GMT (15kb)
Date (revised v2): Tue, 5 Jan 1999 21:36:42 GMT (15kb)
Date (revised v3): Mon, 10 May 1999 04:45:54 GMT (15kb)

String Junctions and Their Duals in Heterotic String Theory

Author: Yosuke Imamura
Comments: 13 pages + 4 eps figures, PTPTeX, typographical errors corrected
Report-no: YITP-99-1
Journal-ref: Prog.Theor.Phys. 101 (1999) 1155-1164

We explicitly give the correspondence between spectra of heterotic string theory compactified on T^2 and string junctions in type IIB theory compactified on S^2.

Author search

Q from submitter: Why should I maintain a bibliography when arXiv will do it for me?

- Examples:
  - http://physics.syr.edu/~bowick/
  - http://astrosun.tn.cornell.edu/faculty/flanagan/flanagan.html (anecdote: third homepage I looked at in CU physics)

Metadata creep?

- arXiv has grown with new technologies and with user expectations extra metadata:
  - Version control data
  - Type information (PS, ignore, encrypted, notebook)
  - Proxy information
  - Recently, DOI
- Some awkwardness in semantics and processing (special cases)

Reformatting and cleaning

- In 2000 we converted all the “old format” records to the current format (14000 records, a combination of automated heuristics and yes/no/edit manual inspection; used cheap/free labor) (Thomas Krichel)
- Tidied all records: removed bad field names, bad dates, checked parenthesis in authors, junk text (e.g. “Abstract: blah“) in abstracts, some TeX junk removed.

Ingest (now)

- No more direct user-entry of metadata in mail-header format. All entry now in web forms.
- Several checks can reject entry (e.g. must be a valid year in a Journal-ref)
- A number of automatic checks flag potential problems to administrators.
- Manual inspection of all new metadata by student administrators.
- Problems also flagged by moderators.

Shortcomings

- Don't insist on affiliation information.
- Can't handle multi-word surnames.
- No way to enter/upload long author lists.
- No linking from Journal-refs.
- Subject classification data free-form, not validated.
- Obscure rules for full-text type and processing metadata
- No way to associate URLs with Journal-ref etc.
- No support for math markup except as TeX
- Where to put copyright statements (many issues)

Authorship

- Author One, Author Two, Author Three
- Author One (Institution 1), Author Two, Author Three (Institution 2)
- Author One (1), Author Two (1 and 2), Author Three (2) ((1) Institution One, (2) Institution Two)

- Currently accept all three forms, plan to demand affiliations, store in structure something like form 3.
- Issues: long lists, collaborations, “for the”, “appendix by”
- Add “role”?

New arXiv metadata format

- Use nested structure (e.g. authorship)
- XML encoding / schema validation
- Use restricted subset of Unicode (native in XML, UTF-8 encoding).
- Influenced by APS experience of ex-arXiv-admin Mark Doyle.
- Still debating level of formatting support (particularly math, subscript etc.).
- Not adopting any existing standard but keeping an eye to cross-walking.

Motivation: Author search

- Popular (link from author name; list of publications)
- Use last name and first initial (bias towards false positive)
- arXiv has multicultural scope, different naming conventions. Focus on notion of "key" name for searching.
  - The surname in our culture
  - May be multi-word (common in Spanish)
  - Some people have one name (India)
  - Some people have initials after surname/key-name
  - (Names using non-Latin scripts?)

Name authority

- The "real" solution to author identification
  - User db could double as name authority db
  - Can't demand that every author register (what about 50 or 1000 author papers!)
- Adopt partial solution?
  - Registered users associate their author name with db record
  - Other authors can later `claim' association
  - How to combine information in search?
- Are we/users ready for this? Extra steps at submission time or short-cuts for registered users?

Motivation: Subject classification

- Currently collect ACM-class for cs submissions; MSC-class for math submissions (1991 or 2000 version?).
- Have lots of bad data (no ingest validation cleaning problem)
- Don't yet collect classification info for physics (PACS codes widely used)
- Adopt model of specifying classification (version) and values
  - <classification type="MSC1991"> <value>33F05</value><value>11F20</value> </classification>
  - Can do some validation of values (…old data?)
- Also working on automated techniques, offer variety

Conflicts

- Ease/convenience for users
- Likelihood of user understanding/correct use
- One third of submitters are new users
- Appropriate granularity and validation
- Use of generalized concepts that work over corpus
- Asking for time investment to learn arXiv system

Non-ASCII characters

- Current metadata format uses ASCII (7-bit)
- Non-ASCII encoded using TeX escape sequences, e.g. \'a for á (render correctly on web)
- Users in original subject areas know TeX, not the case for all subject areas
- New user db uses Unicode, "pidgin TeX" or ISO-Latin input (Paul Houle)

Math markup

- Title and abstract at least
- Adopt some standard (MathML), some subset, or roll our own (copy APS?)
- How will users enter it?
- How will we render it?
- What about existing TeX markup? (Some work by Paul Houle)
- Immature display technology, might be best to wait. Perhaps accept pidgin form now, perhaps do partial TeX translation

Internal bookkeeping

- Information about submitter (links to record in user db)
- Information about other versions of article
- Include size and hash (MD5) for source package, some verification of source fixity
- Extend type to “document model”?
- Probably keep copy of old metadata in <attic> so that it will be there to debug any problems

Interoperability

- Metadata export via OAI (on-the-fly cross-walk to simple DC)
- Export all Journal-ref data to SLAC (daily)
- Import Journal-ref data from SLAC (weekly, automatic)
- Import DOI data from Elsevier (weekly, automatic)
- OpenURL links from arXiv
- Link to citation services: SLAC, Citebase (arXiv id based)

That's all folks…