Effingham Agreement

Thomas Krichel

2000–01–23

0. This text is known as the Effingham agreement. It is available at http://openlib.org/acmes/plan/effingham.html. It is also available as a Portable Document File for A4 and US letter formats. It will take effect when a version has been approved by the the WoPEc steering committee. The other parties to this agreement are Victor M. Lyapunov and Tatyana I. Yakovleva of the Siberian Branch of the Russian Academy of Sciences.

1. arXiv.org, lead by Paul Ginsparg, has been leading efforts to collect preprints in Physics and related areas since 1991. The metadata for the collection has two parsing difficulties. The first parsing difficulty "problem 1" is a combined title and author field in records of a format used between 1991 and 1994 "format 1". This makes it difficult to separate author names from titles of papers. The second parsing difficulty "problem 2" is the combined author and affiliation field that makes it difficult to separate authors from their affiliation. This appears in the format adopted in the format adopted after 1994 "format 2".

2. The work to conduct the ReDIF conversion for the UPS protoproto has been successful up to 90% for "problem 1" and up to about 98% for "problem 2". Since the conversion algorithm is not aware about when a separation has been successful, all records that have "problem 1" have been excluded from the protoproto.

3. Thomas Krichel wishes to further the collection of high quality metadata on freely available scientific document. The plan is to build a ReDIF dataset based on arXiv.org and other data. The data can then be used in user services. It can also be augmented by an author and institution registration service. In order to include the data in "format 1", a correction of "problem 1" is necessary.

4. This agreement concerns the correction of "problem 1". This is to be done through translating metadata files from "format 1" to "format 2". It is hoped that arXiv.org will then use the corrected data instead of the currently held data. This agreement does not cover the "problem 2".

5. Victor M. Lyapunov has a copy the abstract files from arXiv.org on a site local to him. He will then develop a script *lyak_1* that reproduces the unseparated title/author field based on the source abstract metadata, alongside with automatically generated separation.

6. This in an example of this first step

```
astro-ph/9206002
SOURCE: Primordial Nucleosynthesis and the Abundances of Beryllium and
 Boron David Thomas, David N. Schramm, Keith A. Olive, Brian D. Fields
 Plain TeX, 28 pages, 8 figures (not included, but available from
 authors). UMN-TH-1020/92
Title: Primordial Nucleosynthesis and the Abundances of Beryllium and
 Boron David Thomas,
Authors: David N. Schramm, Keith A. Olive, Brian D. Fields
```

In this example we have

- the abstract identifier,
- the source unseparated field, and
- *lyak-1*-generated Title and Authors fields (wrong, in this example).

7. Once these files are written out, they will be transfered to openlib.org. Then Tatyana I. Yakovleva will apply a separator on the data. The separator is chosen to ensure that it does not otherwise appear in the data. For the example above

```
astro-ph/9206002
SOURCE: Primordial Nucleosynthesis and the Abundances of Beryllium and
 Boron David Thomas, David N. Schramm, Keith A. Olive, Brian D. Fields
 Plain TeX, 28 pages, 8 figures (not included, but available from
 authors). UMN-TH-1020/92
Title: Primordial Nucleosynthesis and the Abundances of Beryllium and
 Boron ~David Thomas,
Authors: David N. Schramm, Keith A. Olive, Brian D. Fields
```

8. After that process of manual correction is completed, the corrected records are transfered to openlib.org. Victor M. Lyapunov will then write *lyak-2*. This script processes the manually corrected metadata and constructs the "format 2" source abstract. *Lyak-2* takes the source abstract text and manually tested/corrected *lyak-1* output, and replaces the unseparated Title/Author field with the Title and Author fields, if the correction is done, or with the *lyak-1* proposed separation, if the latter happens to be correct.

9. There is no time limit to the manual separation. It is expected that the *lyak* scripts will take a week each to write and apply.

10. Tatyana I. Yakovleva will receive £.65 per manual correction. The currently expected expense is £910. An advance of £500 will be paid out when the agreement comes into force. The remainder is paid when the "format-2" abstracts are delivered. However, the WoPEc project will not pay more than £1300 for the manual separation.

11. The "format 2" corrected abstract files will be made available to arXiv.org. It is hoped that arXiv.org will use these corrected abstracts to replace the originals.

12. In case of a disagreement on this agreement, the parties appoint Sune Karlsson as a mediator.