DSpace Character Encoding HOWTO

> ***Note: This HOWTO is a work-in-progress -- please contribute if you have any comments or advice to add, or corrections to make!***
>
> *Note: This HOWTO is called "Character Encoding" but mainly deals with enabling Unicode/UTF-8 for repositories wishing to correctly handle and display UTF-8 characters*

Character encoding in DSpace

Character encoding is an important consideration in digital repositories, archives and catalogues. Even if the majority of your digital resources are described in English, or in characters from the ISO-8859-1 (Latin1) character set, it is likely that users will eventually wish to search using characters from scientific character sets, character sets outside ISO-8859-1 (Latin1) or that your repository needs to be compatible with other institutional systems that only speak UTF-8.

This HOWTO will give some tips and tricks to ensure your DSpace repository, user interfaces and servlet container are consistent in their handling of character encoding (and, better yet, compliant with UTF-8). It will also hopefully serve to remind developers of common pitfalls, so they can be avoided in future ;-)

In DSpace 1.5.2 and 1.6.0, many character encoding fixes were submitted to help DSpace become more compliant with UTF-8. Previous versions may find that handling of text in search forms, license text, collection and community names is inconsistent, particularly in XMLUI (Manakin). A list of relevant JIRA issues can be found at the end of this page to help you identify any possible character encoding issues with your version of DSpace.

Character encoding in dspace-api

Character encoding in the core API is often overlooked, but the following components can be affected by incorrect character encoding:

- Emails sent to users and administrators from DSpace
- Log file entries and error messages
- Output from admin tasks like media filters

Character encoding in JSPUI

Character encoding in XMLUI (Manakin)

Character encoding in Apache Tomcat

Character encoding in PostgreSQL

Still to add: Guides for other servlet containers (Glassfish, JBoss, Jetty, etc.); guide for Oracle; other DSpace webapps (SWORD, LNI, OAI, etc.)

Useful links

Reading materials / references

- Tomcat character encoding FAQ
- w3schools' HTML URI encoding reference
- Java internationalisation FAQ

Related JIRA issues

- CLONE - Foreign characters broken in group names.
- UTF-8 encoding in community and collection text
- Special characters in collection license lead to parse error
- Fix configurable browse parameter encoding (XMLUI)
- Scandinavian characters break in license, group names and collection/community metadata
- XMLUI Browse by Author doesn't work for names with special characters (for example: é, è, ö, etc.)
- XMLUI overall UTF-8 encoding is inconsistent and forms do not use UTF-8
- Imporper display of Umlauts / Encoding of messages_de.xml - ID: 2413800

Still to add: Related mailing list threads?