— FAUstairs Blue Note* —
The FAUstairs Glossary Extraction and Curation Process

Michael Kohlhase
Computer Science, FAU Erlangen-Nürnberg
http://kwarc.info/kohlhase

November 25, 2025

Abstract

We describe the initial ideas for the FAUstairs glossary extraction and curation process and the workflows and tooling we envision to supports it.

This blue note is (supposed to be) a living document that describes the current state of the discussion, to serve as an implementation guide and initial documentation for the GloX tool ecosystem.

Contents

*Inspired by the "blue book" in Alan Bundy's group at the University of Edinburgh, FAUstairs blue notes, are documents used for fixing and discussing ϵ-baked ideas in projects by the FAUstairs group (see http://kwarc.info). Unless specified otherwise, they are for project-internal discussions only. Please only distribute outside the FAUstairs group after consultation with the author.

1 Introduction

A Central part of the FAUstairs project ("Formative Assessment for Universities: Strategic Application of Innovative Methods to Raise Study Success Rates" see `https://faustairs.fau.de`) is the development and curation of a **domain model** – i.e. a set of key concepts and their definitions – for large portions of the courses at FAU (and the development of added-value services on top of that to establish formative assessment).

In the following we describe the information sources, the glossary extraction and curation workflows and the GloX tool ecosystem.

2 Information Sources and Stakeholders

The main sources of information for the FAUstairs domain model are the following[1]: EdN:1

1. the module descriptions in Campo: `https://campo.fau.de`
2. the course infrastructure and curriculum data on StudOn: `https://studon.fau.de`
3. the course materials of the instructors.

The first two are available via the DIP system, a centralized infrastructure and data store for synchronization of the FAU learning administration systems provided by the FAU RRZE.

The stakeholders in the GloX process are[2] EdN:2

1. The **degree programs** represented by the program directors (the faculty member formally in charge), the program coordinators and maybe the study advisors.
2. The **departments** that host the degree programs, represented by their speakers and the department manager.
3. The instructors of the mandatory courses of a degree program; here we include the persons who organize the tutorials, homework assignments, and (summative) assessments.
4. The **FAUstairs GloXers** – three pairs of knowledge representation and domain specialists tasked with the GloX process.

3 Workflow

The GloX workflow will consist of two large steps glossary extraction and glossary curation, which we will sketch out in the following:

[1] EdNote: MK: I am sure there are more, need to extend

[2] EdNote: MK: there must be more; extend

3.1 Glossary Extraction

In this step we examine the information sources from section 2 for glossary-relevant information and export it into a curation format (most probably FloDown).

The relevant steps are

1. **Concept Identification**: The domain specialists identify the key concepts in the information source
2. **Concept Annotation**: The concepts are annotated with
   (a) a **symbol** name (a system identifier), the concept in the source serves as the default verbalization.
   (b) (optionally) known **synonyms**, and
   (c) a **definition** (rigorous) or **concept documentation** (less rigorous description); both may be annotated by term references.
3. **Translation**[3]: Where the scientific discourse is international, the concept names are standardized to their English versions. EdN:3

3.2 Domain Model Curation

In this step we collect all the available glossaries, aggregate them into a coherent domain model. The relevant steps are

1. **Collection**: The glossaries are collected and systematically organized into a modular collection, most probably managed and served by MathHub.info.
2. **Annotation**: The definitions are further annotated with term references into the joint domain model by the GloXers.
3. **Aggregation**: this is mainly a de-duplication step, which identifies possible duplicate concepts (probably by their definitions and/or usage patterns).
4. **Canonicalization**: The domain model is compared against the disciplinary learning ontologies, etc.

4 The GloX Tool Ecosystem

We will start off the GloX process with a glossary extraction tool for HTML module descriptions (we assume that we can get them as HTML from the DIP). We will have to process hundreds of module descriptions over the course of the FAUstairs project, so a good workflow support for the three would be helpful.

[3] EdNote: MK: do we want to do this?; I think it will be necessary at least for Math, INF and the natural sciences

4.1 HTML GloX

For concept identification (see above), the user selects a text region with a (verbalization of) a key concept. For concept annotation, GloX provides interaction window that allows to enter the necessary data. GloX should probably include Wikipedia or LLM-based suggestions for the synonyms and definitions, possibly a concept spotter as well. Also, we will need to include a snify-like workflow for the annotating the definitions with term references.

It would be good if the HTML GloX were not restricted to the module descriptions, but could be used for arbitrary HTML documents; so that we could use it via pandoc for other formats as well.

4.2 Other GloXes

Many of the other information sources (see section 2) are not (born as) HTML. It will be critical to infiltrate the native workflows of the respective stakeholders to lighten the workload (and possibly create alternative value). Other formats information sources include

1. LaTeX: here the sTeX format [sTeX] is appropriate, and well-established.
2. MS PowerPoint: here we can build on CPoint [Koh08] to provide GloX functionality; in fact we are starting a re-implementation in a Master's project, which should specialize in GloX functionality initially.
3. MS Word: Here we can build on WOIDE (Word OMDoc IDE; see [AK24; Adr25]).
4. Text and MarkDown: here we can build on FloDown[4] EdN:4

All of the tools export FTML (FlexiFormal Text Markup Language)[5], which is at the heart of the KWARC toolchain. EdN:5

4.3 Development Model

As always, I would like the GloX tool ecosystem to be developed in an iterative fashion: first draft functionality soon, so that we can try and criticize, and advanced functionality/features later. And of course, having a first mockup to show/demo to the program coordinators (they need to have an idea what is coming) soon would be great.

[4] EdNote: MK: how to cite it?
[5] EdNote: MK: how to cite it?

References

[Adr25] Aurelius Adrian. "Integrating Semantic Authoring within Rich-Text Environments into the OMDoc Semantic Ecosystem – WOIDE II". B.Sc. Thesis. FAU Erlangen-Nürnberg, Sept. 2025. URL: https://gl.kwarc.info/supervision/BSc-archive/blob/master/2025/AdrianAurelius.pdf.

[AK24] Aurelius Adrian and Michael Kohlhase. "WOIDE: Semantic Annotation in MS Word — Scaling Mathematical User Interfaces beyond LaTeX". In: *MathUI 2024: The 15th Workshop on Mathematical User Interfaces*. Ed. by Kazuhisa Nakasho and Jan Frederik Schaefer. 2024. URL: https://kwarc.info/kohlhase/papers/mathui24-woide.pdf.

[Koh08] Andrea Kohlhase. "Semantic Interaction Design: Composing Knowledge with CPoint". PhD thesis. Computer Science, Universität Bremen, Apr. 2008. URL: https://kwarc.info/ako/pubs/AKo_Promo.pdf.

[sTeX] *sTeX: A semantic Extension of TeX/LaTeX*. URL: https://github.com/sLaTeX/sTeX (visited on 05/11/2020).