— 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).
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

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