BT TOOL VALUE CHAIN / **CROSS-CHAIN THREADS AND STANDARDS**

TSWG
ETEN Tech Strategy Work Group
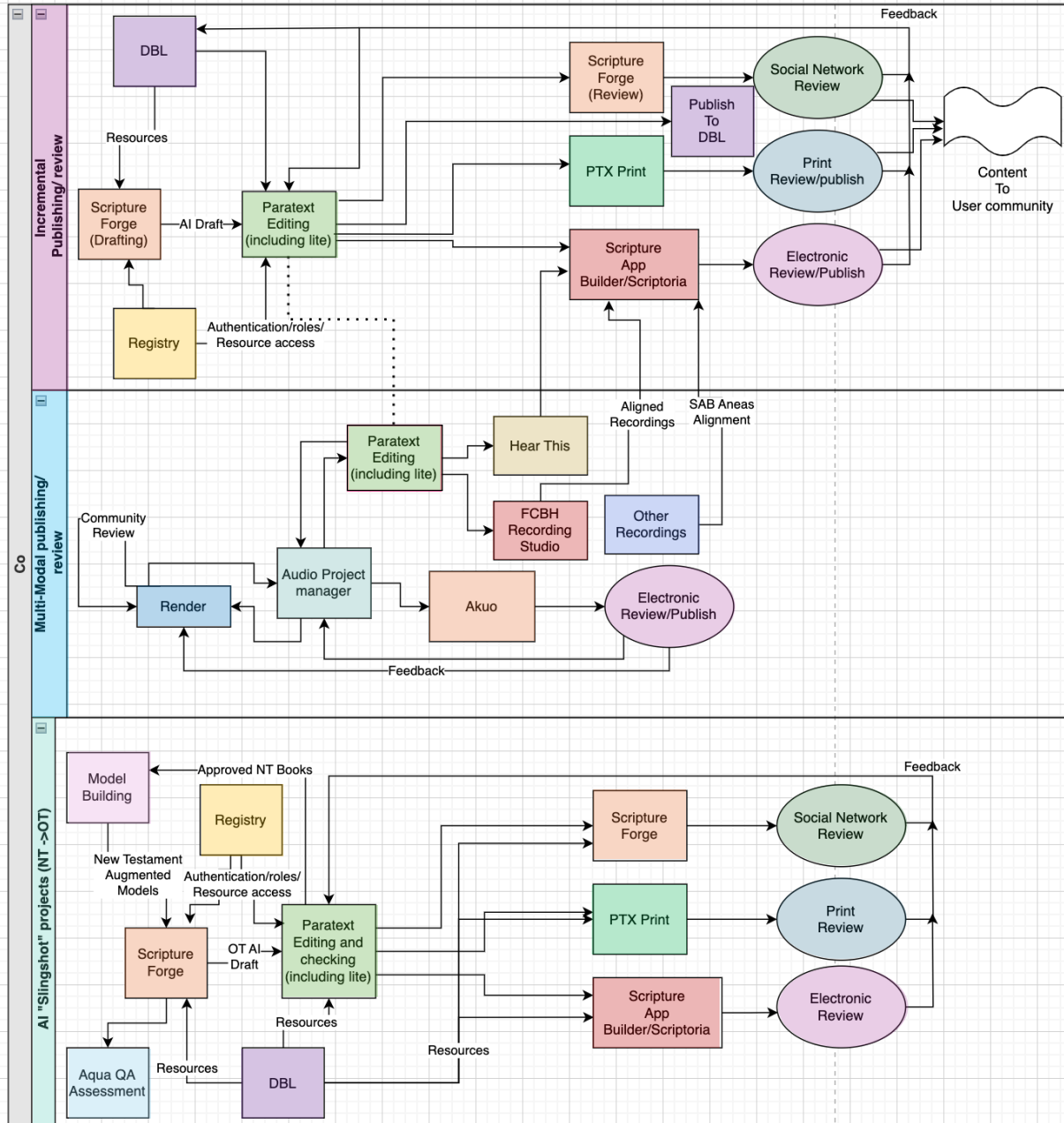
STANDARDS:

Language Technology developers are actively engaged in developing data standards for language data. In collaboration with others in the industry, progress is being made on the following standards:

LIFT — LIFT (Lexicon Interchange FormaT) is an XML format for lexical information (dictionaries). LIFT allows movement of data between programs such as WeSay, FLEx and Lexique Pro.

FlexText — an XML format for interlinear data, to allow movement of data between programs such as SayMore, FLEx and ELAN.

USX - Unified Scripture XML (USX) is an XML format used for encoding the digital text for scripture translations. The largest collection of USX encoded scripture is currently found within the Digital Bible Library.

DBL Text Bundle - XML bundle that contains both the USX content and metadata to capture publishable Scripture. Paratext provides the uploader client used for submitting scripture text content to DBL. The scripture text itself is maintained within Paratext by the translation or text maintenance team. The Paratext project includes different sections of configuration settings which supply a portion of the metadata required for submission to DBL. The Global Bible Catalogue (GBC) provides the remainder of the required metadata. Paratext validates the scripture text and the gathered metadata as conforming to the standards and syntax defined for a DBL 'text bundle'.

USFM - Unified Standard Format Markers (USFM) is a plain text markup widely used for encoding the digital text for Scripture translations. It is the standard format applied to translations developed within ParaTExt

Unicode — an industry-wide character set encoding standard designed to support the worldwide interchange, processing, and display of the written texts of the diverse languages and technical disciplines of the modern world. It is closely related to ISO/IEC 10646.

ISO 15924 — the International Organization for Standardization's registry of scripts. Each script is identified by a name and four-letter code. The current version of the standard includes 156 scripts.

ISO 639-3 — the International Organization for Standardization's registry of the languages of the world. It is comprised of living languages taken from SIL's Ethnologue, as well as extinct, ancient, reconstructed, and artificial languages. The current registry includes over 7,000 languages, each identified by a unique three-letter code.

Scripture Burrito – Standards for Bible translation tool interoperability.