Zulip Terminology Stream Text Mining Project Description

Base Zulip bulletin board application is supported by a REST API that can be interrogated via Python scripts. Bots can be configured via Python to provide real-time monitoring as well. Text mining of the terminology stream in the FHIR Zulip community bulletin board to discover trends regarding use of terminologies and terminology services within the HL7 FHIR community.

Objective of this exercise is to review the history of the content and activity Terminology stream for the following:

1) Identify Code System (names) in topics, also in response / body text
   a) DICOM
   b) SNOMED
   c) LOINC
   d) ICD*
   e) NDC
   f) RxNorm
   g) (.... a few others…)

   Count the total numbers of these (rank order)
   Note dates of first / last activities within each thread

2) Identify HL7 Product families other than FHIR in topics, also in response / body text.
   Presumes thread deals with some interaction between product family and FHIR
   a) V2 / Version 2
   b) CDA or C-CDA
   c) V3 / or Version 3

   Count the total numbers of these (rank order)
   Note dates of first / last activities within each thread

3) Identify FHIR Terminology Resources in topics, also in response / body text
   a) ConceptMap*
   b) CodeSystem*
   c) ValueSet*
   d) Terminology Service*
   e) TerminologyCapabilities*
   f) NamingSystem*
   g) Code*
   h) Coding*
   i) CodeableConcept*

   Count the total numbers of these (rank order)
4) Identify FHIR Terminology Operations in topics, also in response / body text
   a) $lookup
   b) $validate-code
   c) $subsumes
   d) $find-matches
   e) $expand
   f) $validate-code
   g) $translate
   h) $closure

   Count the total numbers of these (rank order)
   Note dates of first / last activities within each thread

5) Count indicators of JIRA activity - ie: links to HL7 JIRA pages, PSS’, tickets etc.
   Count the total numbers of these (rank order)
   Note dates of first / last activities within each thread

6) Age / date analyses.
   a) Determine average length (in days / wells / months) in terminology stream threads.
   b) Identify outliers in length - longer running threads
   c) Identify those topics with many responses (not necessarily with longer length, but will likely be one of these as well) that do not have some sort of resolution. Will require iterative review with SME (Davera or others)
   d) For each of the count categories (1-4) above, when is the occurrence of these topics, when are they more frequent / less frequent
   e) Date-base counts for all topics indicating activity levels: when is the stream more active / less active