Challenges in Deploying and Managing Large Terminologies: NCI Thesaurus

For Protégé Workshop
June 22, 2009
Amsterdam

Gilberto Fragoso, Sherri de Coronado

Challenge Topics

- Background
  - EVS, NCI Thesaurus, Vocabulary Development, Distribution Methods
- Core Challenges
  - Content, many users and uses
  - Support, Properties, Provenance
  - Editing Tool Requirements
    - Shared Data and Distributed Editing
    - Simplified GUI, Reasoning, Rule Enforcement, etc.
- Other Issues/ Challenges
  - Training and Maintaining pool of domain expert editors
  - Modeling Consistency
  - QA
  - Collecting Input from end users and collaborators

Background

- NCI Enterprise Vocabulary Services
  - Support
  - Products
- Distribution
  - LexEVS Terminology Server
  - Web Browsers (Bioportal, NCI Thesaurus and Metathesaurus)
  - FTP site (OWL and other formats)
- Production Cycle

NCI Enterprise Vocabulary Services

Goal – Integration by Meaning

- EVS provides services and resources that assists to:
  - Integrate different conceptual frameworks for clinical, basic and translational research,
  - Create terminological and taxonomic conventions across systems
- Controlled Terminology Products
  - NCI Thesaurus – an ontology-like cancer-centric controlled terminology
  - NCI Metathesaurus – maps biomedical vocabularies
  - External vocabularies maintained and served: MedDRA, HL7, NDF-RT, LOINC, GO, Zebrafish, RadLex, etc.
  - BiomedGT (Biomedical Grid Terminology - new)
- Further info, see: https://wiki.nci.nih.gov/display/EVS/EVS+Wiki

Vocabulary Support Guidelines

- Enable appropriate use of multiple terminologies and mappings between them.
- Leverage existing sources where appropriate
  - VA NDF-RT, RxNorm, LOINC, etc. …
  - Develop unique content where needed (Cancer genes and diagnoses, drugs and therapies, molecular abnormalities, clinical trial standard terminology etc.)
- Link to other information sources and standards using URLs as possible
  - GO, Swissprot, drug formularies, trial protocols etc.
- Merge with or map as needed to other standard terminology to ensure interoperability

Products: NCI Thesaurus

- **Reference Terminology for NCI, caBIG, Partners**
  - Underpins caCORE, caGRID semantics
- **A Federal Standard Terminology**
- About **80,000 "Concepts"** hierarchically organized into domains
- **Broad coverage** of the cancer research and clinical domain including prevention and treatment trials
  - Neoplastic and other Diseases
  - Findings and Abnormalities
  - Anatomy, Tissues, Subcellular Structures
  - Agents, Drugs, Chemicals
  - Genes, Gene Products, Biological Processes
  - Animal Models – Mouse, other
  - Research techniques and management, apparatus, clinical and lab, radiology, imagery
- **Published Monthly**

Products: NCI Thesaurus (2)

- **Public domain**, open content license
- **Description-logic based**
- **Concept History**
- **Distributed in multiple ways**:
  - By **download** (OWL, Ontylog XML, flat files)
  - Through LexEVS 3.2 (in deprecation), LexEVS 4.2 and **LexEVS 5.0 server** and **caGRid terminology node**
  - As a source in **NCI Metathesarus** and UMLS Metathesaurus
  - Online Via **Browsers**
    - NCI Bioportal: http://bioportal.nci.nih.gov/ncbo/faces/index.xhtml
    - Brand New: http://ncit.nci.nih.gov (NCIt specific browser)

Distribution: LexEVS

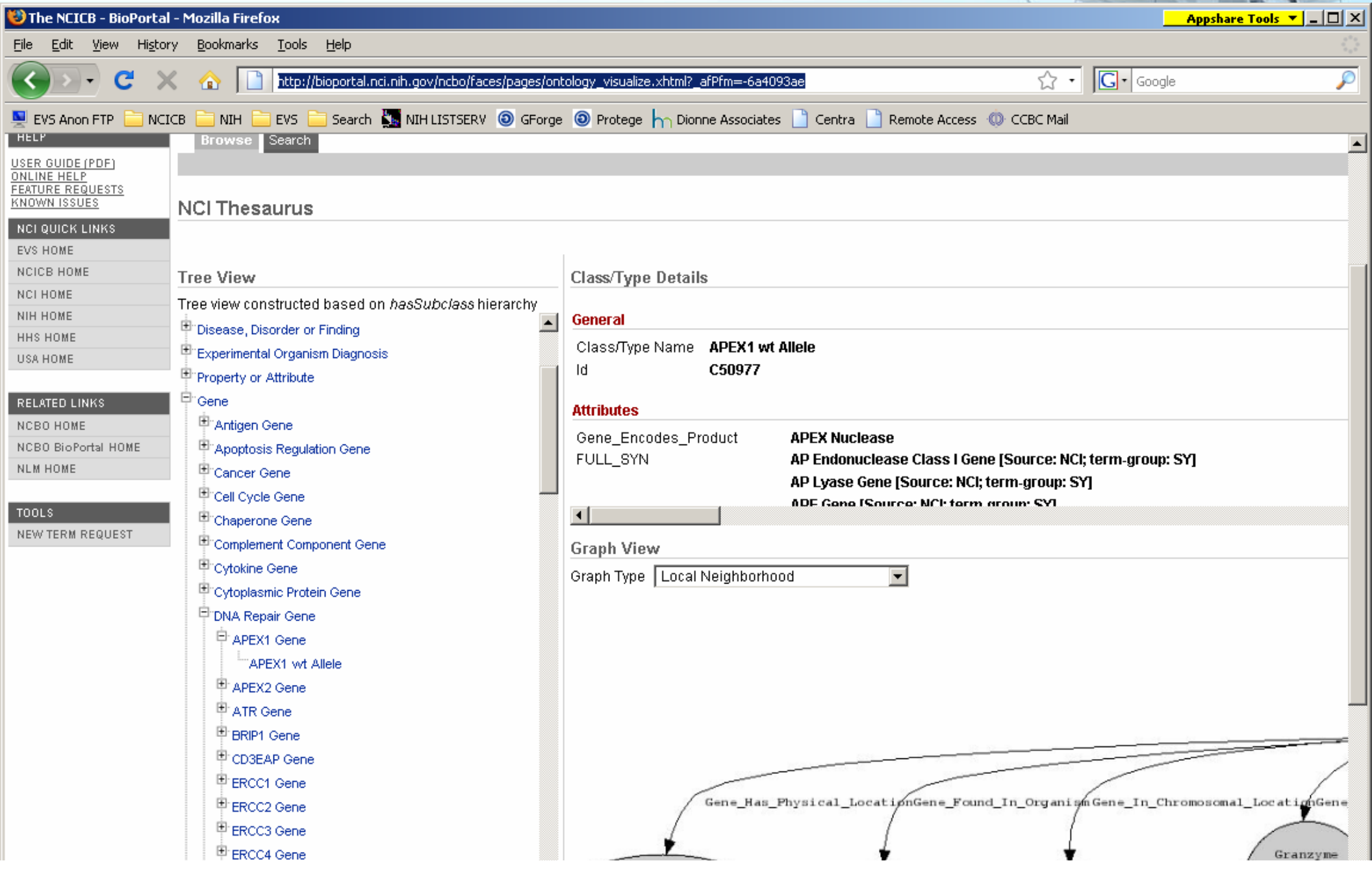
- What is LexEVS?
  - LexEVS is a collection of APIs that provide access to controlled terminologies.
  - The controlled terminologies hosted by the NCI EVS Project are published via the Open-Source LexEVS Terminology Server.

Distribution: LexEVS 5.0

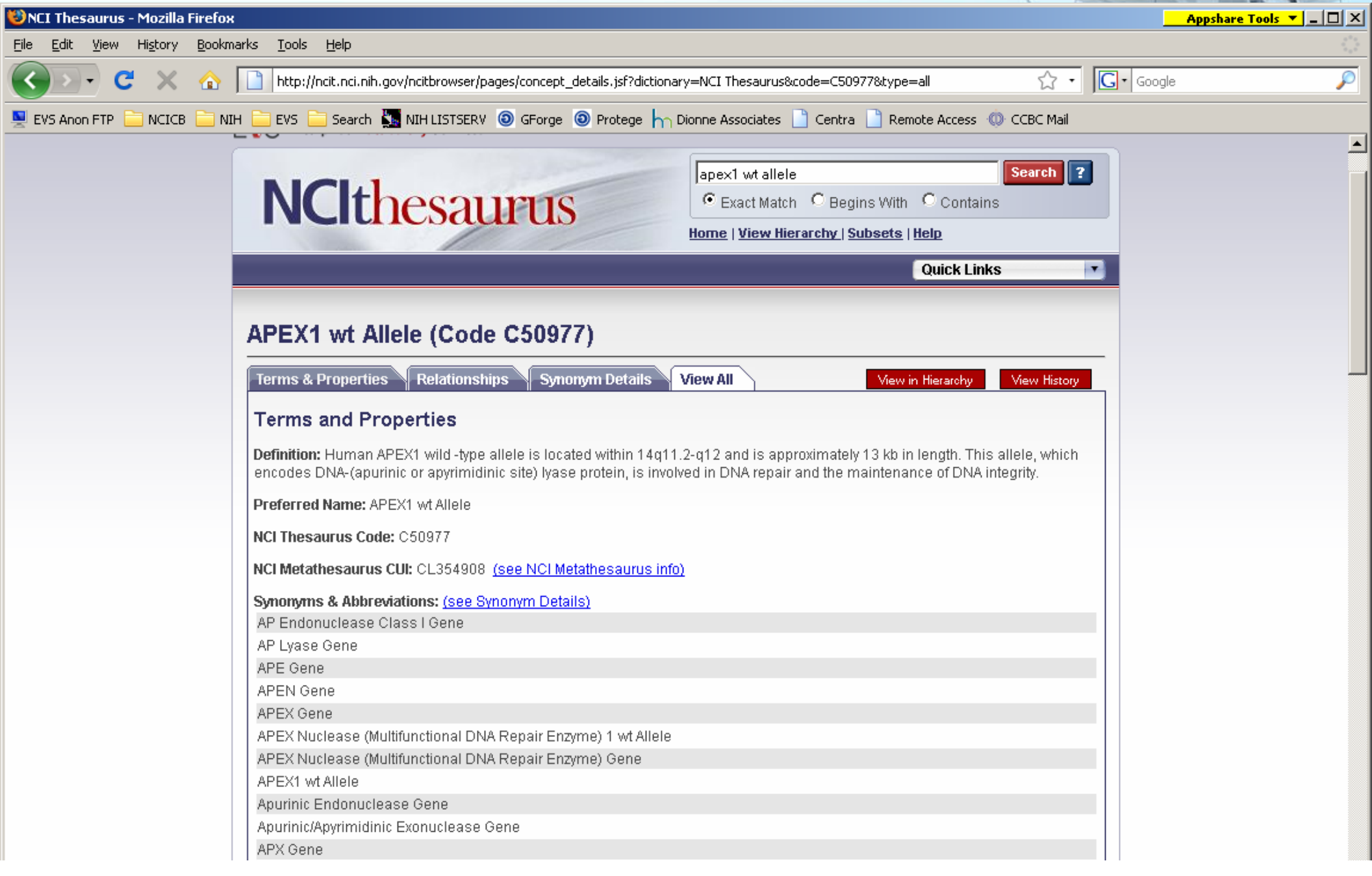
The LexEVS 5.0 Release includes the following components:

- Java API - A Java interface based on the LexGrid 5.0 Object Model
- REST/HTTP Interface - Offers an HTTP based query mechanism. Results are returned in either XML or HTML formats
- SOAP/Web Services Interface - Provides a programming language neutral Service-Oriented Architecture (SOA)
- Distributed LexBIG (DLB) API - A Java interface based on the LexGrid 2009/01 data model and relies on a LexEVS Proxy and Distributed LexEVS Adapter to provide remote clients access to the native LexEVS API
- LexEVS 5.0 Grid Service - An interface which uses the caGRID infrastructure to provide access to the native LexEVS API via the caGRID Services
- See: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexBig_and_LexEVS for information and

NCI Thesaurus in NCI Bioportal



NCI Thesaurus Browser



EVS Products & Services Are Open

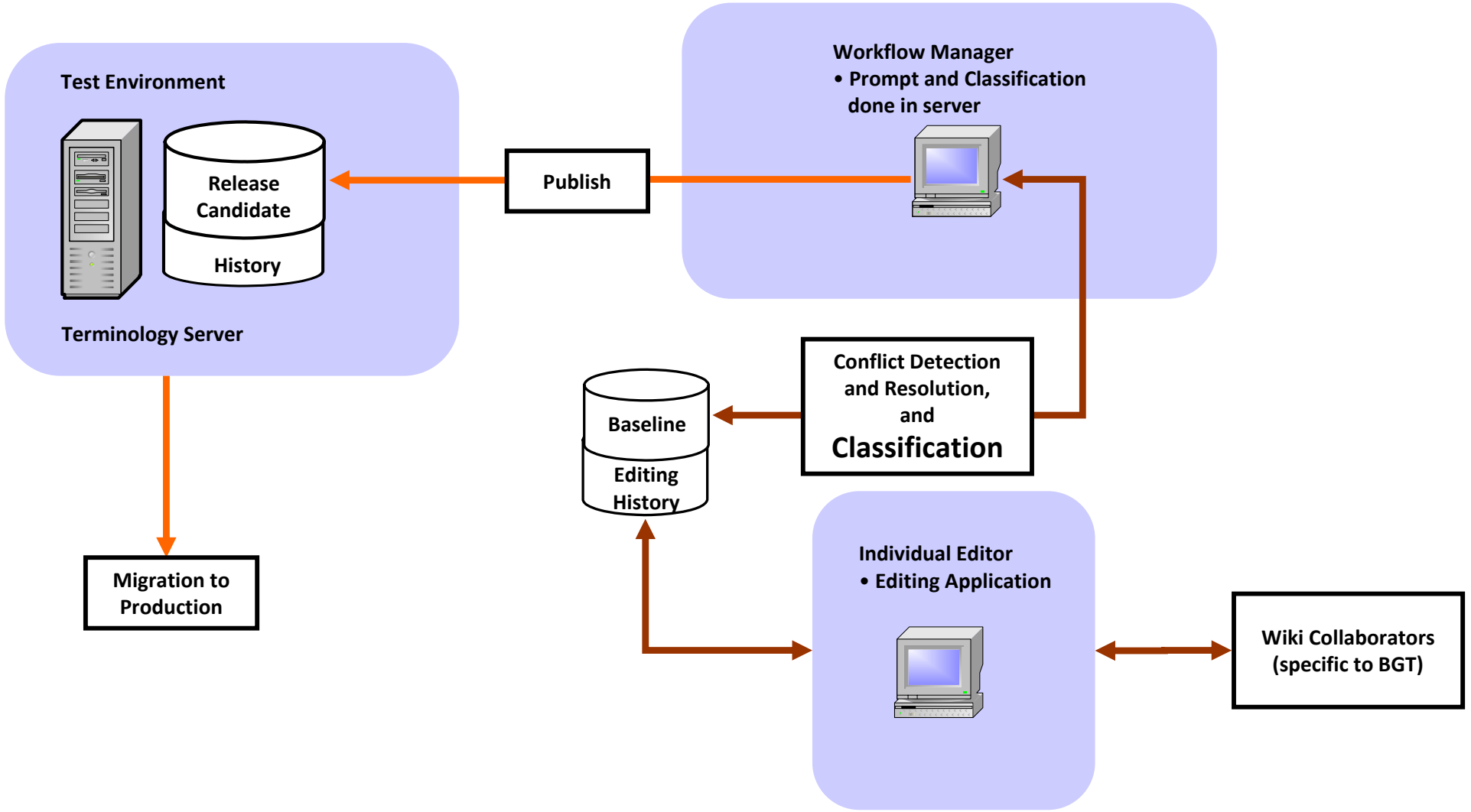
- NCI Thesaurus and BiomedGT Are Open Content
  ftp://ftp1.nci.nih.gov/pub/cacore/EVS/NCI_Thesaurus/ThesaurusTermsofUse.htm
- NCI Metathesaurus is Mostly Open Source

  See Each Source's License
  http://ncimeta.nci.nih.gov/MetaServlet/GenerateSourcesServlet
- NCI EVS Servers Are Freely Accessible
  - ***On the* Web:**

  http://ncit.nci.nih.gov, http://bioportal.nci.nih.gov/ncbo/faces/index.xhtml, *and* http://ncimeta.nci.nih.gov
  - ***Via* API or caGRID:** See: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexEVS_5.0_Components Read Me file for API urls.
- All Software Developed by NCI EVS is Public Open Source:

  http://ncicb.nci.nih.gov/download/cacoreevsapilicenseagreement.jsp

Current NCIT and BGT Production Environment



Core Challenges

- Content: many domains, users and uses
- Support, Use of Properties, Tracking Provenance
- QA and Editing Consistency
- Editing Tool Requirements
  - Shared Data and Distributed Editing
  - Simplified GUI
  - Editing Consistency
  - Reasoning,
  - Rule Enforcement, etc.

Many NCIt Users and Uses

- Content Challenges: Wide variety of users being supported simultaneously :
  - NCIt provides the foundation for semantics in caBIG (NCI and partners), used by caDSR and by applications to annotate metadata and data
  - Used by FDA and CDISC, to develop and distribute terminology subsets for Structured Product Labels, Study Data Tabulation Model, etc.
  - Used as standalone by number of applications

Additional Use Cases

- Coding and Data (Drug / Clinical) Integration
  - Agents, Clinical Trials and Adverse Events
    - CTEP and DCP clinical trials, unambiguous identifiers
    - PDQ Cancer Clinical Trials Registry & NCI Drug Dictionary
    - Federal Medication Terminologies (FMT)
    - FDA Structured Product Labeling, e.g. pill shape
- Semantic Interoperability in caBIG
  - caTIES/caTissueCore/caMOD/caNanolab
- Harmonization (CDISC/ FDA/ BRIDG/ ISO DT)
- We don’t know all the users!

Newer Use Cases

- Query and reasoning against instance data on the Grid
- Federation of ontologies and subontologies (BiomedGT)

Challenge: Supporting Different Requirements with Annotations

- Annotations used to record concept info:
  - Provenance (history tracking, contributors)
  - Lexical information (terms, definitions)
  - Support of external programs (vocab subsets)
  - Authoritative information (e.g. OMIM, NSC)
  - Usage (scope notes)
- Challenge: standard terminology?
  - alt_term, synonym, full_syn
  - definition, def, comments
  - SKOS gaining traction, but lacking in some areas (provenance)

Challenge: Editing Consistency

- Modeling consistency
- Description Logic

  used to construct better hierarchies
- Editor Guide
- Design Guide
- Programmatic support to enforce edit checks and business rules
- QA performed at various stages in the production cycle

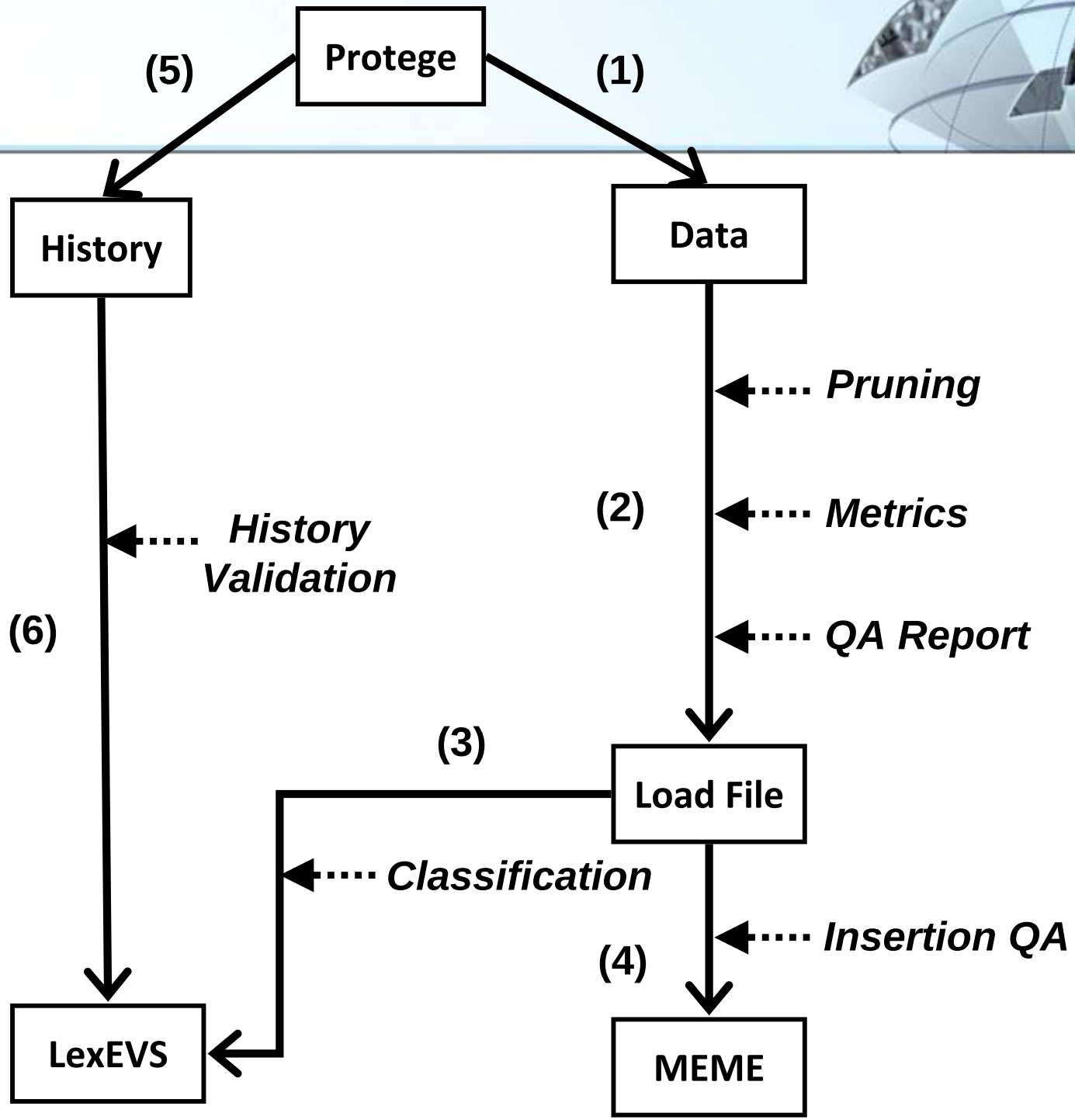
Quality Assurance

- Combination of Manual and Automated Processes
  - Consistency checking with reasoner
  - Edit checks built into software
  - Editing and Design documents reviewed and updated periodically
  - Edit checks built into production cycle
  - Internal (ongoing) and External (Periodic) reviews
- See: Journal of Biomedical Informatics 42 (2009) 530–539. The NCI Thesaurus quality assurance life cycle Sherri de Coronado, Lawrence W. Wright, Gilberto Fragoso, Margaret W. Haber, Elizabeth A. Hahn-Dantona, Francis W. Hartel, Sharon L. Quan, Tracy Safran, Nicole Thomas, Lori Whiteman

Edit Checks Configured into SW

Table 1 – Selected edit checks built or configured into the software

Edit Check Entity	Description
Concept Name	Cannot be changed (although preferred term can be); it must begin with letter or underscore.
Preferred Name	Concept must have one and only one Preferred Name; it must match the fully qualified synonym with group PT and source NCI.
Duplicates	Duplicate parents, roles and properties are not allowed.
Definition	Each must have 1 review date, 1 review name, 0 or 1 attributes; no characters less than utf-8 32 allowed; !,? or @ allowed, single spaces only in definitions unless preceded by these special characters.
Retired Concept	Only lead editor can retire concepts, although editors can pre-retire concepts. An editor's note should explain the retirement.
Merged Concept	Only lead editor can merge concepts, although editors can pre-merge concepts, and pre-merged concept must include an editor note with value of pre-merge annotation and an explanation if needed.
Split Concept	Check that newly created concept is a valid concept. All checks made during normal create are made during a split.
Other	Cannot create or maintain a restriction relationship that points at a retired, pre-retired or pre-merged class.


Protege
(5)
(1)
History
Data
Pruning
(2)
Metrics
History
Validation
QA Report
(6)
(3)
Load File
Classification
Insertion QA
(4)
LexEVS
MEME

QA Steps During Processing

Table 2: QA check steps during history processing

Check Description	Sample output (condensed)
Write a log file to characterize edits as create, merge, retire, modify for concept history file.	674642|C73624|create|30-APR-08|(null) 674643|C38019|split|30-APR-08|C38019 674644|C38019|split|30-APR-08|C73624 674659|C3279|modify|30-APR-08|(null) 674661|C72063|modify|30-APR-08|(null)
Check for concepts that have appeared but have no create record.	(No error example found.)
Check for concepts that have disappeared.	Concepts found in C:\...\TDEByNameForProduction-07.05e.xml but not in C:\...\TDEByNameForProduction-07.06d.xml C67256
Check for history records that don't have matching concepts.	New concepts not found in BSLN2 (C:\...\TDEByNameForProduction-08.08d.xml): C75530|Hyperimmune_State
Check for invalid merge codes.	(no examples, caught by edit filters)
Check for concepts created and retired within an editing period.	WARNING: New codes created, then retired, but still found in BSLN2:(to be edited manually) C75602|Motion 687348|C75602|Motion|New|2008-08-22 04:08:12.0|<editor etc>
Multiple modifies of a concept for period combined into 1 history record.	List of all discarded records: 687355|C75604|IDS_Gene|Modify|2008-08-25 10:08:27.0|<editor etc>
Editor identity information removed from records.	687347|C2558|Glufanide_Disodium|Modify|2008-08-22 02:08:04.0|<editor etc> 687348|C75602|Motion|New|2008-08-22 04:08:12.0|<editor etc> 687349|C75603|Artifact|New|2008-08-22 04:08:20.0|<editor etc> 687347|C2558|modify|12-SEP=08|(null) 687348|C75602|create|12-SEP=08|(null) 687349|C75603|create|12-SEP=08|(null)
Discard modify records on new concepts.	Modify records correponding to new codes are discarded: 687355|C75604|IDS_Gene|Modify|2008-08-25 10:08:27.0|<editor etc> 687358|C75604|IDS_Gene|Modify|2008-08-25 10:08:48.0|<editor etc>
Discard modify records on merged concepts.	Modify records correponding to merged codes are discarded: 688366|C15721|Epidemiology_Research|Modify|2008-09-03 09:09:21.0|<editor etc> 688367|C71483|Epidemiologic_Study|Modify|2008-09-03 09:09:23.0|<editor etc>

Editing Tool Requirements

- **Shared Data and Distributed Editing**
- **Reasoning**
- **GUI for Domain Experts (not ontologists)**
- **Editing Consistency**
  - **basic content – preferred and alternative terms, definition**
  - **number and types of restrictions needed**
- **Complex Operations**
  - **merge**
  - **split**
  - **retirement**
- **Rule Enforcement & Guidance**
  - **no duplicate restrictions**
  - **definition, semantic type**

Other Issues/ Challenges

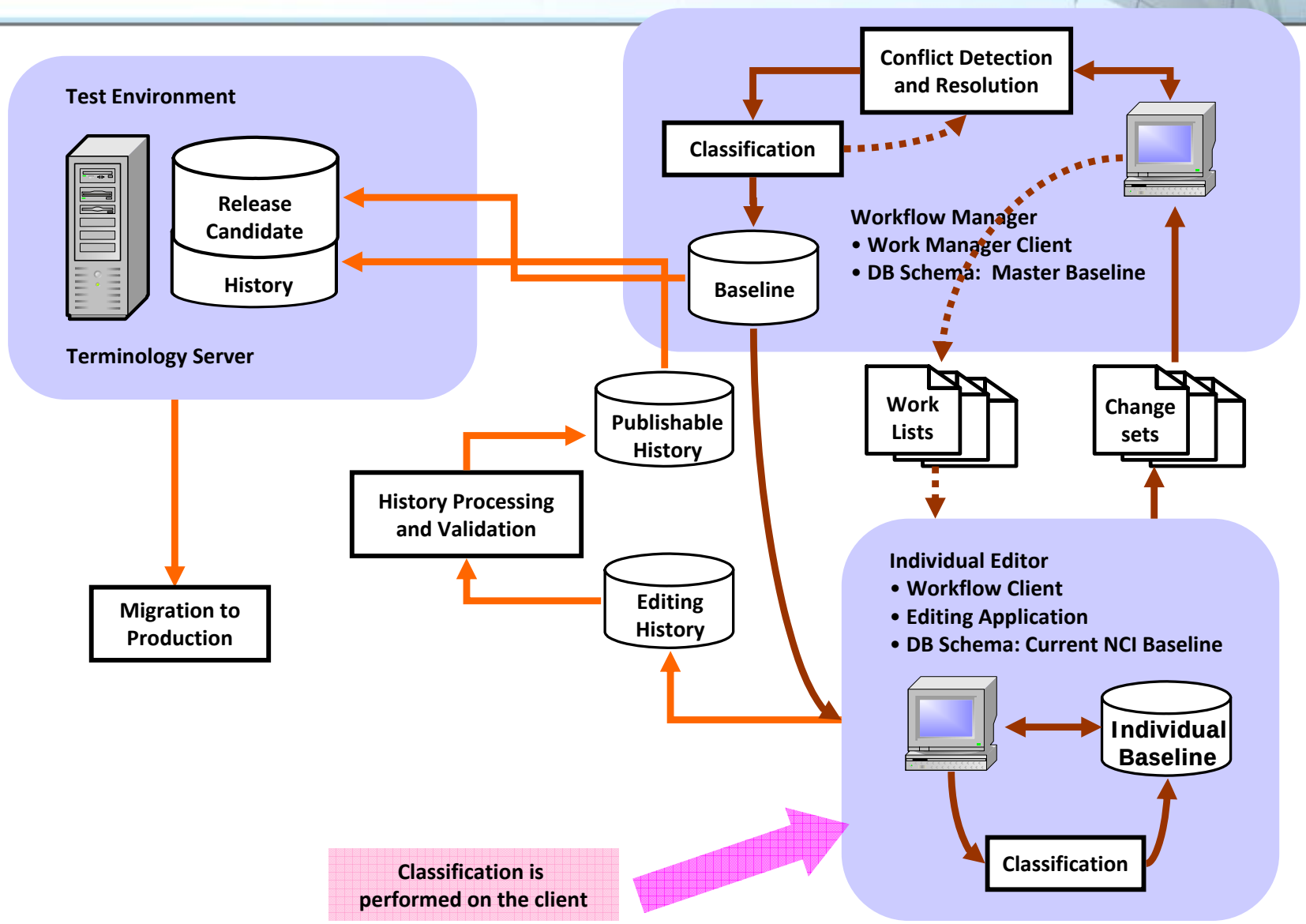
- **Training and Maintaining pool of domain expert editors**
- **Modeling Consistency**
  - **Description Logic**
    - **Use to catch errors in model/modeling**
  - **URU properties**
  - **However, guidelines and consensus are still necessary**
    - **are concepts modeled "fully" or**
    - **are concepts modeled just enough to make them defined**
- **QA**
  - **We modify QA process as new issues arise**
- **Collecting Input from end users and collaborators**
  - **Not everybody is an ontologist, simple interfaces are necessary, allow domain experts to work on what they know**

Acknowledgements

EVS Team

- NCI staff
  - Frank Hartel
  - Gilberto Fragoso
  - Sherri de Coronado
  - Margaret Haber
  - Larry Wright
- Editing and QA
  - Laura Roth, Lori Whiteman, Liz Dantona (mangers) +10 others (Lockheed Martin)
- Terminology Servers
  - Mayo, Northrup-Grumman
- Production and QA staff
  - Tracy Safran (SAIC)
  - Rob Wynne,
  - Sharon Quan et al (LM-Alameda)
- Protégé/ NCI Protégé programmers
  - Stanford BMIR staff
  - Dionne Associates
  - Northrup Grumman
  - Clark & Parsia

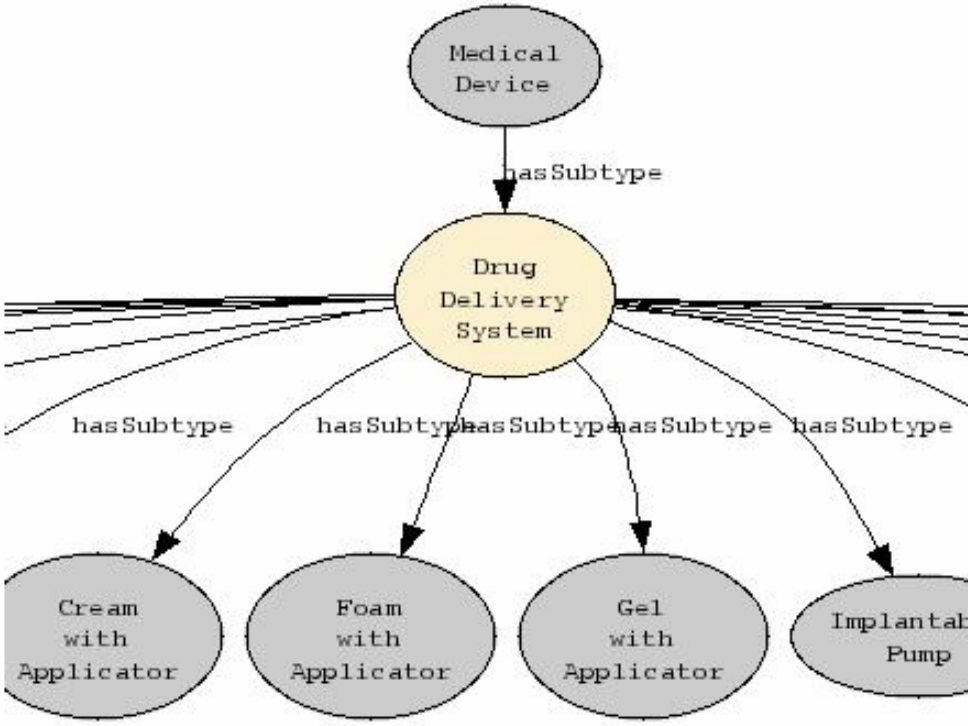
NCIT Production Environment

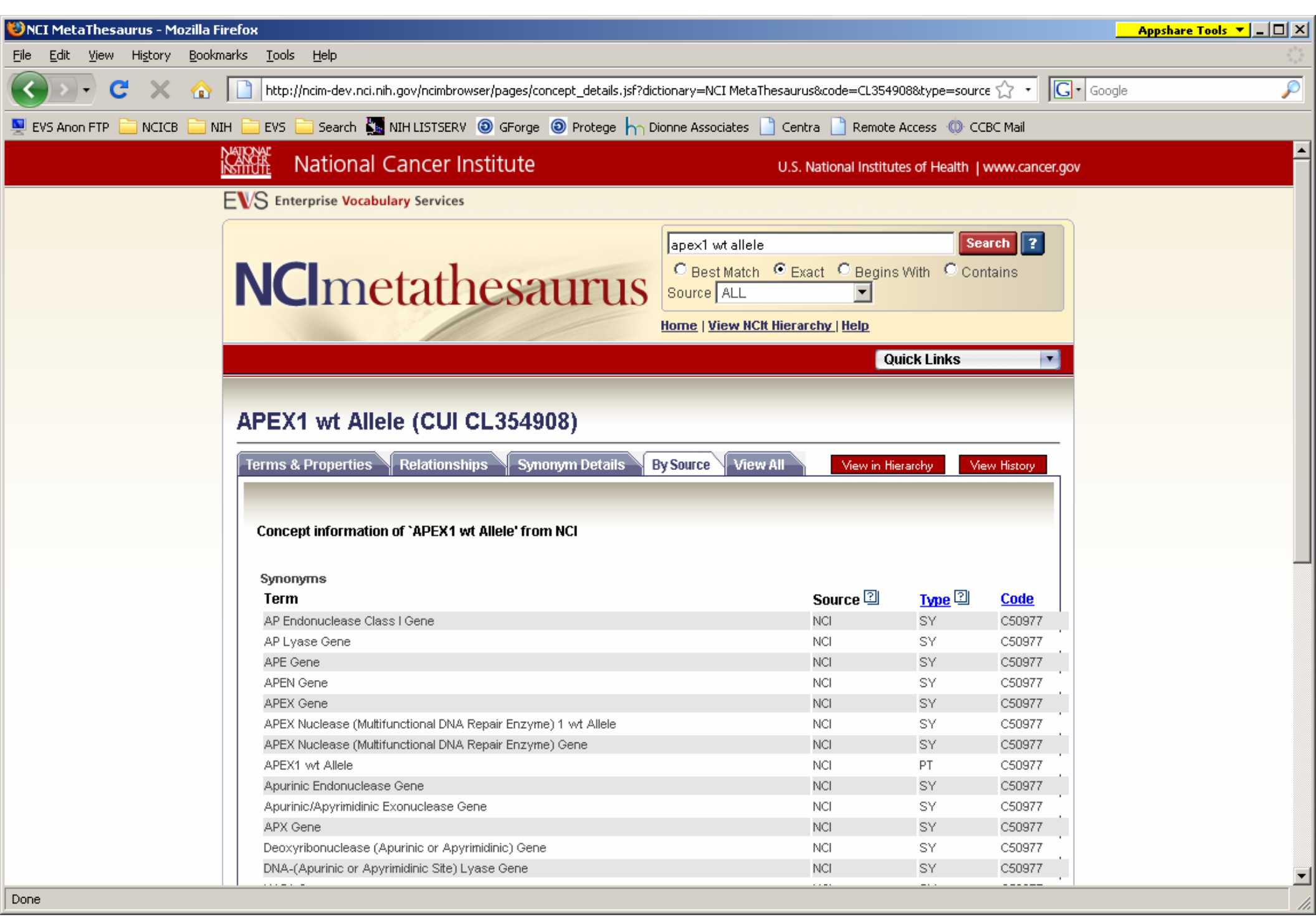


Delivery System,
Drug Delivery
Systems

Graph View

Graph Type: Local Neighborhood




NCI MetaThesaurus - Mozilla Firefox
Appshare Tools
File Edit View History Bookmarks Tools Help
http://ncim-dev.nci.nih.gov/ncimbrowser/pages/concept_details.jsf?dictionary=NCI MetaThesaurus&code=CL354908&type=source
Google
EVS Anon FTP NCICB NIH EVS Search NIH LISTSERV GForge Protege Dionne Associates Centra Remote Access CCBC Mail
National Cancer Institute
U.S. National Institutes of Health | www.cancer.gov
EVS Enterprise Vocabulary Services
NCImetathesaurus
apex1 wt allele
Search
Best Match Exact Begins With Contains
Source ALL
Home | View NCIt Hierarchy | Help
Quick Links
APEX1 wt Allele (CUI CL354908)
Terms & Properties Relationships Synonym Details By Source View All
View in Hierarchy View History
Concept information of `APEX1 wt Allele' from NCI
Synonyms
Term Source Type Code
AP Endonuclease Class I Gene NCI SY C50977
AP Lyase Gene NCI SY C50977
APE Gene NCI SY C50977
APEN Gene NCI SY C50977
APEX Gene NCI SY C50977
APEX Nuclease (Multifunctional DNA Repair Enzyme) 1 wt Allele NCI SY C50977
APEX Nuclease (Multifunctional DNA Repair Enzyme) Gene NCI SY C50977
APEX1 wt Allele NCI PT C50977
Apurinic Endonuclease Gene NCI SY C50977
Apurinic/Apyrimidinic Exonuclease Gene NCI SY C50977
APX Gene NCI SY C50977
Deoxyribonuclease (Apurinic or Apyrimidinic) Gene NCI SY C50977
DNA-(Apurinic or Apyrimidinic Site) Lyase Gene NCI SY C50977
Done

NATIONAL CANCER INSTITUTE

EVS ENTERPRISE VOCABULARY SERVER

Basic Advanced

apex1 wt allele Search

The basic search is enabled. Click on the "Advanced" tab to customize your search criteria.
You can mouse-over each advanced search item for help in utilizing the advanced features.

About
Browse
Copyright
New Term
Sources
User's Guide

Concept | Definitions | Synonyms | Sources | Broader Concepts | Narrower Concepts | Related Concepts | Medications | Procedures | Laboratory | Diagnosis | Open NCI Hierarchy | View Hierarchy Location

CL354908: APEX1 wt Allele

Gene or Genome

APEX1 wt Allele Definitions

Source	Definition
NCI2009_02D	Human APEX1 wild -type allele is located within 14q11.2-q12 and is approximately 13 kb in length. This allele, which encodes DNA-(apurinic or apyrimidinic site) lyase protein, is involved in DNA repair and the maintenance of DNA integrity.

APEX1 wt Allele Synonyms

- APEX1 wt Allele
- Human Apurinic Endonuclease 1 Gene
- Redox Factor 1 Gene
- REF-1 Gene
- AP Endonuclease Class I Gene
- AP Lyase Gene
- APE Gene
- APEN Gene
- APEX Nuclease (Multifunctional DNA Repair Enzyme) Gene

Center for Bioinformatics

NCIt: Example Concept (1 of 2)

Preferred Name: Gastric Mucosa-Associated Lymphoid Tissue Lymphoma

Code: C5266

Semantic Type: Neoplastic Process

Parent Concepts: Extranodal Marginal Zone B-Cell Lymphoma of Mucosa-Associated Lymphoid Tissue

Gastric Non-Hodgkin's Lymphoma

Synonyms & Abbreviations: Gastric MALT Lymphoma

Gastric MALToma

(subset) MALT Lymphoma of the Stomach

MALToma of the Stomach

Primary Gastric MALT Lymphoma

Primary Gastric B-Cell MALT Lymphoma

Primary MALT Lymphoma of the Stomach

Definition: A low grade, indolent B-cell lymphoma, usually associated with Helicobacter Pylori infection. Morphologically it is characterized by a dense mucosal atypical lymphocytic (centrocyte-like cell) infiltrate with often prominent lymphoepithelial lesions and plasmacytic differentiation. Approximately 40% of gastric MALT lymphomas carry the t(11;18)(q21;q21). Such cases are resistant to Helicobacter Pylori therapy.

NCIt: Example Concept (2 of 2)

Role Relationships (subset) for Gastric Mucosa-Associated Lymphoid Tissue Lymphoma:

Molecular abnormalities:

Disease_May_Have_Cytogenetic_Abnormality: Trisomy 3
Disease_May_Have_Cytogenetic_Abnormality: Trisomy 18
Role group 1:
Disease_May_Have_Cytogenetic_Abnormality: t(11;18)(q21;q21)
Disease_May_Have_Molecular_Abnormality: AP12-MLT Fusion Protein Expression

Histogenesis:

Disease_Has_Normal_Cell_Origin: Post-Germinal Center Marginal Zone B-Lymphocyte

Pathology:

Disease_Has_Abnormal_Cell: Centrocyte-Like Cell
Disease_May_Have_Abnormal_Cell: Neoplastic Monocytoid B-Lymphocyte
Disease_May_Have_Abnormal_Cell: Neoplastic Plasma Cell
Disease_May_Have_Finding: Lymphoepithelial Lesion

Anatomy:

Disease_Has_Primary_Anatomic_Site: Stomach
Disease_Has_Normal_Tissue_Origin: Gut Associated Lymphoid Tissue

Clinical information:

Disease_Has_Finding: Primary Lesion
Disease_May_Have_Finding: Indolent Clinical Course
Disease_May_Have_Associated_Disease: Hepatitis C