NCI Enterprise Vocabulary System

Vocabulary Executive Board

Quarterly Meeting, April 3, 2002

The Role of the EVS in Creating a Public Domain Vocabulary

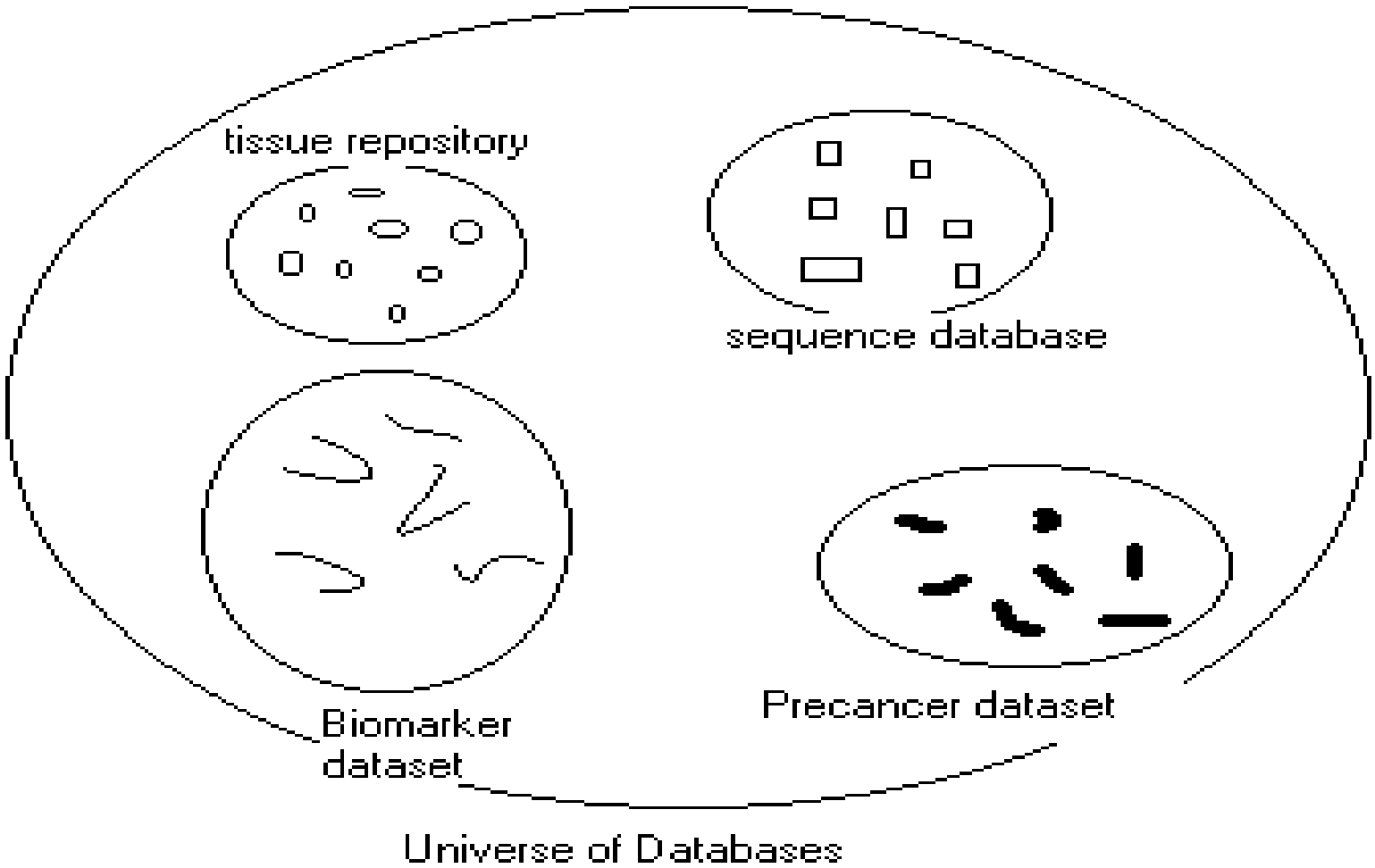
Jules J. Berman, Ph.D., M.D., Cancer Diagnosis Program

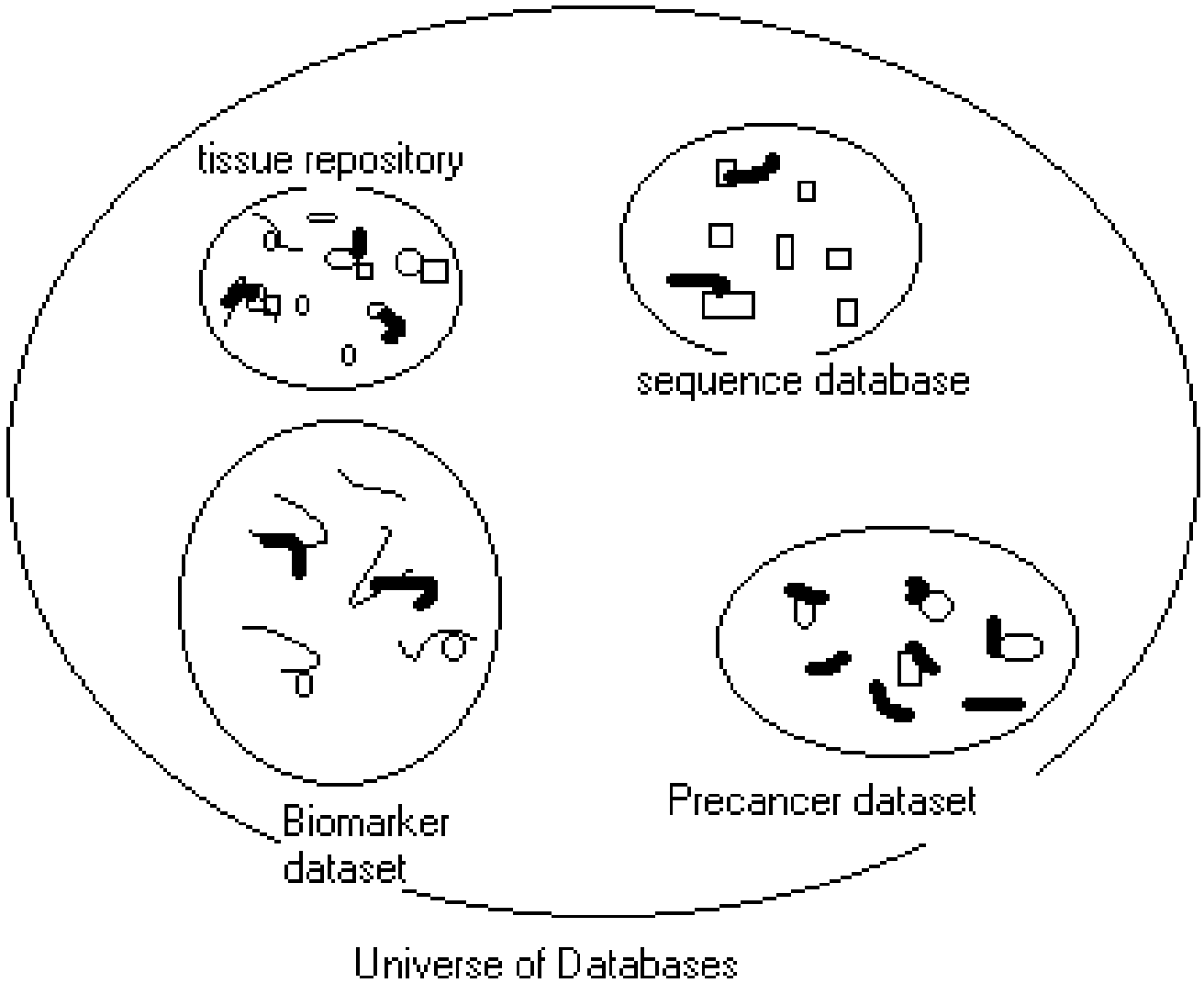
Why do we need a freely available nomenclature?

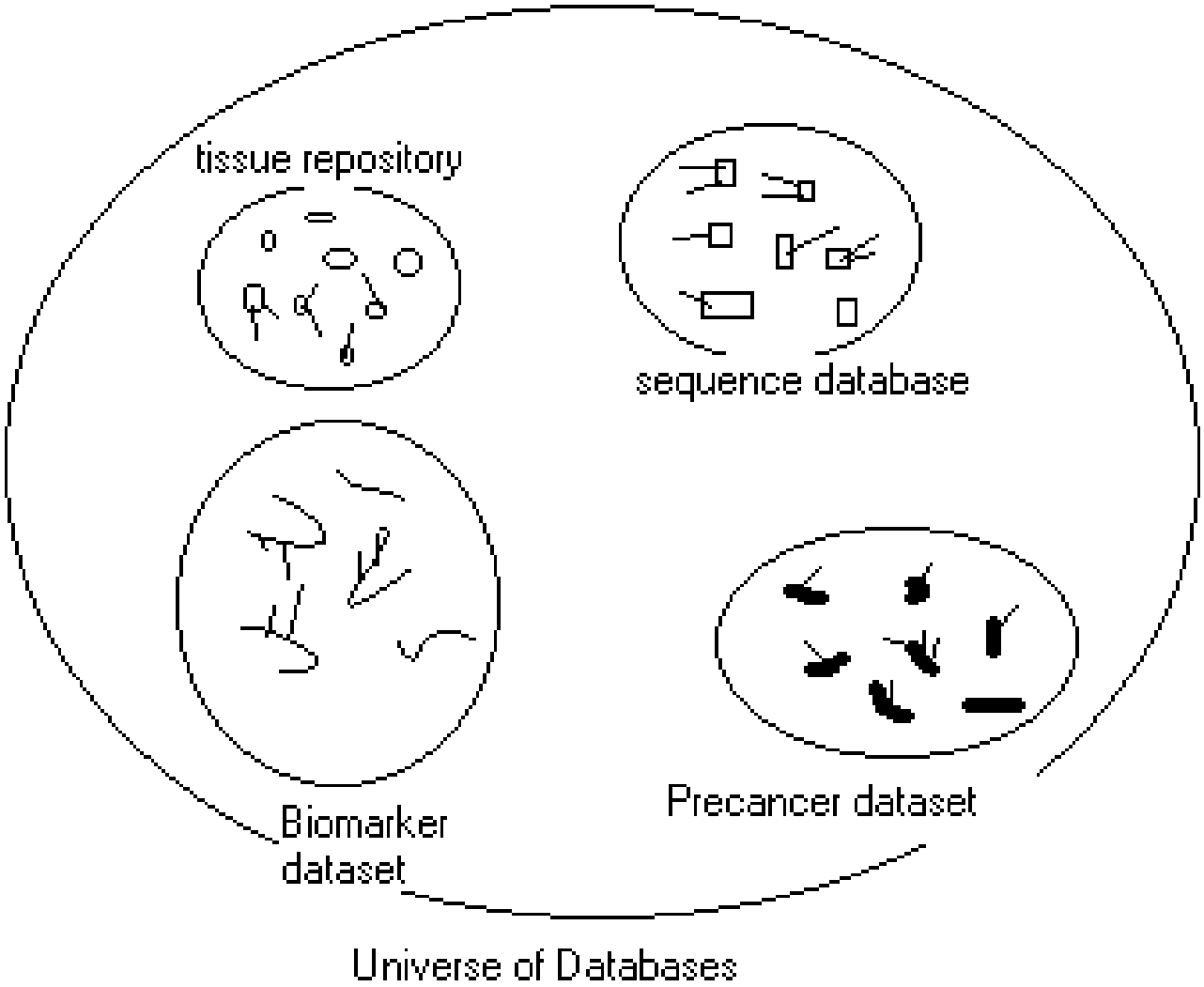
So that medical data can be organized, retrieved and analyzed.

So that bioinformatics data can link to medical informatics data.

It needs to be free so that any researcher in any lab in any country can access research data and contribute to the research process (something that can already be done with genomic data)


tissue repository
sequence database
Biomarker dataset
Precancer dataset
Universe of Databases


tissue repository
sequence database
Biomarker dataset
Precancer dataset
Universe of Databases


tissue repository
sequence database
Biomarker dataset
Precancer dataset
Universe of Databases

Medical data needs to be annotated with standard codes that represent informative medical descriptors.

These medical descriptors occur in free-text, and a given medical concept can be expressed many different ways.
It is necessary to have a way of identifying a given concept regardless of the choice of terminology in the text.

Example:

renal cell carcinoma
renal cell carcinomas
carcinoma, renal cell
hypernephroma
hypernephroid carcinoma
grawitz tumor
renal cell adenocarcinoma
rcc

All map to UMLS CUI C0007134

Wrong: Structured reporting is just around the corner. Everyone will use the same terms for entering data, and we won't need to worry about free text issues.

Ask yourselves: in the past 10 years, how much of your keyboarding has been structured entry and how much has been free-text?

Is the amount of free-text inputting been going up or down?

Wrong: We already have free nomenclatures for medicine, like ICD

UMLS is encumbered by proprietary vocabularies

ICD is not made for computer parsing

MESH may be the closest thing we have to a free annotation vocabulary.

Wrong: Everyone uses SNOMED and has SNOMED licenses, so SNOMED is the de facto nomenclature for the free exchange of medical annotations

Different versions of SNOMED (incompatible with each other)

License doesn't permit electronic transfer of codes

Latest market strategy is to charge for per/case coding. Impossible when you want to encode a large database of millions of legacy reports

Wrong: There may be restrictive language written into the SNOMED licenses but nobody pays attention to the fine print.

I have had personal experiences with two major efforts involving large medical datsets and deep pocket institutions. They take the fine print in the SNOMED license very seriously.

Wrong: Any day now there will be a federal contract with SNOMED that will permit SNOMED to be used freely by everyone.

Don't hold your breath.

How can NIH broker a contract that will permit a researcher in Singapore to have free access to SNOMED-coded data that was freely created by an unfunded researcher in America?

Bottom line: Nobody shares [with the research public] datasets coded with a proprietary vocabulary.

Wrong: It costs money to make a nomenclature, so people have to expect to pay for its use.

That doesn't seem to be much of an issue in the BIOINFORMATICS community.

Human Genome Project is freely available to the world
Taxonomy.dat is freely available to the world
GO (Gene Ontology) is freely available

TAXONOMY .DAT 36,782,346 EBI SITE

//
ID : 725
PARENT ID : 727
RANK : no rank
GC ID : 11
SCIENTIFIC NAME: Haemophilus influenzae biotype aegyptius
SYNONYM : "Bacillus aegyptius" Trevisan 1889
SYNONYM : "Bacillus conjunctivitidis" Kruse 1896
SYNONYM : "Bacterium aegyptiacum" Lehmann and Neumann 1899
SYNONYM : "Bacterium conjunctivitis" (sic) Chester 1897
SYNONYM : "Bacterium pseudo conjunctivitidis" (Kruse 1896) Chester 1
SYNONYM : "Hemophilus conjunctivitidis" (sic) (Kruse 1896) Bergey et
SYNONYM : Bacillus aegyptius
SYNONYM : Haemophilus aegyptius
SYNONYM : Hemophilus conjunctivitidis
SYNONYM : Haemophilus influenzae aegyptius
SYNONYM : Bacillus conjunctivitidis
SYNONYM : Bacterium aegyptiacum

SYNONYM : Bacterium aegyptiacum
SYNONYM : Bacterium conjunctivitis
SYNONYM : Bacterium pseudo conjunctivitidis
SYNONYM : Haemophilus aegyptius (Trevisan 1889) Pittman and Davis 19
SYNONYM : Haemophilus influenzae biogroup aegyptius
COMMON NAME : Koch-Weeks bacillus
MISSPELLING : Haemophilus aegypticus
//

Note: for some medical terms, even common terms, the misspelling occurs more often than the correct spelling or the abbreviation occurs more often than the expansion. Most terminologies miss the issues of misspellings and abbreviations.

What do we need to do?

1. Survey the existing free taxonomies

UMLS concepts unencumbered subset (58 Mbytes)
Gene Ontology (9 Mbytes)
Taxonomy.dat (36 Mbytes)
Omim entities (90 Mbytes)
published lists of abbreviations and misspellings

2. Determine how these can be compiled into a single vocabulary

3. Determine how the vocabulary can be curated and grown

Identify stakeholders? Open Source effort?