

# XML at Genentech -- Not yet?

John “Scooter” Morris, Ph.D.

Computing Technologies

Genentech, Inc.



# Outline

Setting

Impediments to XML

XML Strategy



# Caveats

**I think XML is an important technology**

**I was asked to talk about impediments**

**I asked other IT staff members**

**Most of the following are their perceptions**



# Setting

***“Genentech is a pharmaceutical company dedicated to applying recombinant DNA technologies to unmet medical needs.”***

## Statistics

- ~3,700 Employees
- ~\$1.3B in Revenue
- 8 products
  - Protropin, Nutropin, NutropinAQ, NutropinDepot, Activase, Pulmozyme, Herceptin, Rituxan
- Three major sites
  - South San Francisco, California
  - Vacaville, California
  - Porrino, Spain
  - Several U.S. Sales offices



# Setting - II

## Heavily regulated industry

- US Food and Drug Administration GxP regulations
- Built-in resistance to change
- Build vs. Buy is highly slanted towards buy

## Strong Research Focus

- Very academic environment
- Lots of need for innovation
- Time to market is key
- Build vs. Buy is more balanced: build when advantageous



# Impediments to XML

**Confusion**

**Standards**

**Tools**

**Vendors**



# Impediments to XML - Confusion

## What is XML?

- XML is HTML with the ability to define your own tags
  - Can't use it until browsers support it
- XML is a data-definition language
  - Competes with: UML, CORBA, EJB, etc.
  - Just use Oracle (or Rational) instead
- XML is a way to store unstructured data so that it can be searched and correlated
  - Our consultant said we don't need a database
- XML is a document-interchange language
  - We'll use it when vendors support it

## What is required to use XML?

- Do you need CSS or XSL or a DTD or ...



# Impediments to XML - Standards

## How many standards are too many?

W3C adds to the open domain

### Extensible Markup Language (XML)

[Working Drafts \(Developer Discussion\)](#) · [Events/Pubs \(translations\)](#) · [Software](#) · [Bookmarks](#)

The Extensible Markup Language (XML) is the universal format for structured documents and data on the Web. [XML in 10 points](#) explains XML briefly. The base specifications are [XML 1.0](#), W3C Recommendations Feb '98, and [Namespaces](#), Jan '99. The [XML Activity Statement](#) explains the W3C's work on this topic in more detail. For related work, see:

Healy · [XML Schema](#) · [XPath](#) · [XPath 2.0](#) · [DOM](#) · [RDF](#) · [CSS](#) · [XSL](#) · [JITML](#) · [Web HTML](#) · [SVG](#) · [XML Signature](#)

early

### Releases

- Implementation of draft standards
- Proprietary (or vendor specific) extensions

## Transition to released standards could cause problems





# Impediments to XML - Tools

**No Commercial Tools for DTD development**

**Need more industry-specific DTD's**

- OASIS is a good start
- Biztalk repository seems to compete with OASIS

**Need better XML (and CSS) support in commercial browsers**

**Need more commercial XML editors with tag support**

- Its going to be difficult to use multiple editors
- We already have a standard text editor

**Freeware tools are viewed with some distrust**

- Particularly in non-R&D areas



# Impediments to XML - Vendors

## Strong vendor enthusiasm is good

- Lots of vendors have an XML strategy
- Unfortunately, they focus on different aspects of XML
- Some finger-pointing: “their XML isn’t real”

## XML needs to be core

- Many vendors “also support” XML

## XML has become a buzzword

- Vendor: “We’re going to use XML as an open data exchange standard”
- DNA: “Are you going to publish your DTD?”
- Vendor: “I’m sorry, I’m not familiar with that term”

## Regulatory agencies have not adopted it

- Still operate under the paper-standard (PDF)



# Genentech Strategy

## Teach technical staff

- Need to identify which XML standards are most important to us (to shorten the learning curve)
- Enthusiasm is building

## Push vendors

## Start selected pilot projects

## Work with industry groups and FDA

- Adopt industry specific DTD's:
  - Clinical trial results
  - Adverse event reporting
  - Sequence patent information
  - Assay results



# XML Pilot Projects

## Project management schema

- Internal to Information Resources
- Get an internal success story

## Genbank schema

- Very important in R&D
- Possibility for a very visible “win”

## Manufacturing Data Warehouse

- Major collaboration sponsored by Genentech
- Will encourage use of XML



# Future XML Projects

## Project management schema

- Standard structured document repository
- Want XML Schemas
- Goal: automated capture and categorization based on specific fields



## Project Charter

---

---

**Project Name:** \_\_\_\_\_

Written by:

Date: \_\_/\_\_/\_\_

Status: *Approved or Draft*

Sponsor:

Owner:

### Project Description

#### Stakeholders

Client Sponsor and Owner are identified at the top of the page. List all other stakeholders in the project, including process and policy stakeholders.

#### Project Scope

Document specifically what is in and out of scope.

#### Regulatory Context

Will this system need to be validated? If yes, list which regulations apply.

#### Project Priority

System Required Date:

What is the priority of the project with respect to other IT projects proposed and in-progress?

#### Resources Required/Assigned

Project Manager:

Identify any project resources already known. Determine skills and knowledge needed from the client and the IT community. Estimate cost of acquiring additional resources, including retractable resources (temps, consultants) and new hardware and software purchases.

Use Resources/Skills Requirements Worksheet, if necessary.

#### Maintenance & Support Requirements

##### Complexity

Technical and business complexity of the project, using Complexity Analysis Worksheet, if necessary.

##### Assumptions, Constraints, Issues, Risks

Use Risk Assessment Worksheet, if necessary.

#### Reviews and Approvals

Project Request

Requested by:	Written by:	Approved by:
Date __/__/__	Date __/__/__	Date __/__/__



# Future XML Projects

## Genbank schema

- Public database of genetic data
- Already an ASN.1 definition
- ASCII is still the “standard”
  - Lots of variations
- Data is corporate critical
- Used as input for lots of tools and analysis
- Desire is to have it be Web accessible
- Possible use of Dublin Core for author/publication information



LOCUS HUM4STS170 357 bp DNA STS 11-MAY-1995  
 DEFINITION Human chromosome 4 sequence-tagged site STS4-170, sequence tagged site.  
 ACCESSION L00739  
 VERSION L00739.1 GI:806398  
 KEYWORDS STS; human chromosome 4; sequence tagged site.  
 SOURCE Homo sapiens, clone p4-110 from a chromosome 4 plasmid library; vector pBluescript II KS + (Stratagene).  
 ORGANISM Homo sapiens  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.  
 REFERENCE 1 (bases 1 to 357)  
 AUTHORS Goold,R.D., diSibio,G., Xu,H., Lang,D.B., Dadgar,J., Magrane,G., Dugaiczky,A., Smith,K.A., Cox,D.R., Masters,S.B. and Myers,R.M.  
 TITLE The development of sequence-tagged sites for human chromosome 4  
 JOURNAL Hum. Mol. Genet. 2, 1271-1288 (1993)  
 MEDLINE 94004872  
 COMMENT On May 11, 1995 this sequence version replaced gi:177266.  
 PCR components: 25 ng of human genomic DNA, 10 pmol of each oligonucleotide, 200 micro-M dNTPs, 0.25 U Taq polymerase (Cetus) in 10 micro-l of 50 mM KCl-20 mM Tris-HCl, pH 8.3 (at room temp), 2.5 mM MgCl-2. Initial denaturation at 94degC for 1.5 min, then 30 cycles of 94degC for 15 sec, 62degC for 23 sec, and 72degC for 30 sec, followed by a final extension at 72degC for 3.5 min, using a Perkin-Elmer 9600 thermocycler. PCR-amplified product size 173 bp.  
 Sequence submitted by:  
 Human Genome Mapping Center  
 Box 0925  
 University of California San Francisco  
 San Francisco, CA 94143-0925 USA  
 Phone: (415) 502-1612 Fax: (415) 476-8391  
 e-mail: hgmcprobes@cgl.ucsf.edu.  
 FEATURES Location/Qualifiers  
 source 1..357  
 /organism="Homo sapiens"  
 /db\_xref="taxon:9606"  
 misc\_feature 4..27  
 /label=Primer1  
 misc\_feature complement(152..176)  
 /label=Primer2  
 BASE COUNT 119 a 68 c 49 g 111 t 10 others  
 ORIGIN  
 1 cccatgggtt gtgcatacct aaaaatccca tggatggat ttccaatgtg aaatctggg  
 61 acaactcaat ctgggcaaac acaacaaat gaactacaat taggtcatga agaactcag  
 121 aagaataaga tgttctattt taaaacctac actacttcat tcagtctaaa atccgtcate  
 181 gangttggtta tttatcattc tgcttataat taacataacc acctactgtn gngcoattta  
 241 atctatttnt tcttacttag agtgacaaga ccaactgatta tttggtaatt accannnma  
 301 tgcoatgta tataaaaagt gtatncccaa tcaaaatctt catatcaaat ctgagat





# Future XML Projects

## Manufacturing Data Warehouse

- Lots of data from a variety of sources
- Need to use lots of tools
- Both discreet and continuous data
- Two major collaborators
  - A data warehouse vendor
  - A process control vendor

