A Standard Format as a Knowledge Artifact in Research Communities

Susan Elliott Sim
University of California, Irvine
ses@ics.uci.edu

Overview

- Introduction to GXL
- Emergence of GXL as a Standard
- Participating Communities
  - Software reverse engineering
  - Graph transformation
  - Graph drawing
- Lessons for Evolving Knowledge Artifacts

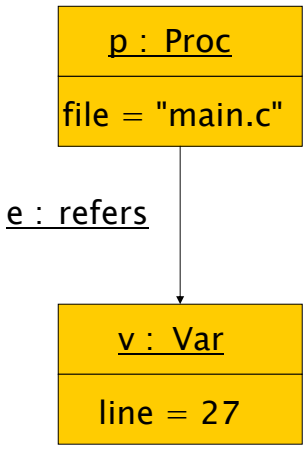
Summary

- Claim: A data format is a knowledge artifact.
- Claim: Successful standardization is a community-driven process.
- Claim: A successful standard data format is a community-driven knowledge artifact.
- Claim: Standardization provides insight into the underlying technical problem and participating scientific communities.

Introduction to GXL

- Graph eXchange Language
- An XML sub-language for exchanging graphs
- Underlying data model is a typed, attributed graph
  - Analogous to databases using tables as underlying data representation
- Schema and instance data are represented separately
  - Analogous to distinction in databases
- Schema and instance data use a uniform representation- same DTD
  - Standard schemas to be defined for various domains

typed attributed graph




```
<?xml version="1.0"?>
<!DOCTYPE gxl SYSTEM "gxl.dtd">
<gxl><graph>
<node id = "p">
  <type xlink:href =
        "schema.gxl#Proc"/>
  <attr name = "file">
    <string>main.c</string></attr>
</node>
<node id = "v">
  <type xlink:href =
        "schema.gxl#Var"/>
  <attr name = "line">
    <int>27</int></attr>
</node>
<edge id = "e"
  from = "p" to = "v">
  <type xlink:href =
        "schema.gxl#refers"/>
</edge>
<graph></gxl>
```

Purpose

- Evolved out of many independent efforts to enable data interoperability
  - Many special-purpose formats
- Format for exchanging data between tools
  - Use the best tool for the job
  - Avoid re-inventing the wheel
  - Facilitate tool interoperability

Nature of Data Formats

- Claim: A data format is a knowledge artifact.
- Reflects current understanding of the problem and state of solutions
- Persistence
  - Statement of what data is important to record
  - Relevance across contexts
- Many implicit assumptions
  - Model of tool and its place in the world
  - Data model/schema

Overview

- Introduction to GXL
- Emergence of GXL as a Standard
- Participating Communities
  - Software reverse engineering
  - Graph transformation
  - Graph drawing
- Lessons for Evolving Knowledge Artifacts

Current Status

- Ratified by reverse engineering community
- Ratified by graph transformation community
- Efforts ongoing in graph drawing
- Recognized graph format outside of these fields
  - Concurrency, model checking, statistical computing, genomics…
- Used by ~40 research groups in 8 countries

- Schema-based data interchange is a current research problem
  - These features have not been used

Design Team

- Andreas Winter, University of Koblenz
  - Visual languages, meta-modeling, graph formalisms, reengineering
- Ric Holt, University of Waterloo
  - Reverse engineering, relational algebra
- Andy Schürr, (now) University of Darmstadt
  - Graph transformation, graph grammars

- Me
  - Empirical studies, program comprehension, pragmatist, communicator

Nature of Standardization

- Claim: Successful standardization is a community-driven process.

- *De facto* vs. *de jure* standards
- Standardization reflects consensus
  - Need for a solution
  - Correctness of solution
- Some group of users agrees and accepts a format, method, benchmark,…

Standard Data Formats

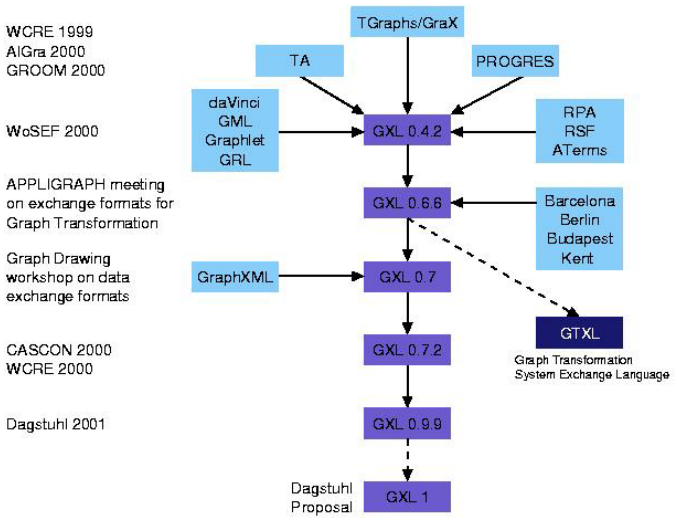
- Claim: Successful standard data formats are community-driven knowledge artifacts.

- Inherits properties from both data formats and standardization process
- A community statement of current state of understanding of a problem and solutions embodied by tools.
- Can't standardize what we don't understand.

GXL Process

- Consensus-based process
  - a.k.a. retail diplomacy
- Led by champions
- Supported by laboratory work
- Lots of opportunities for feedback
  - Meetings at conferences, workshops, seminars
  - Mailing lists, web sites, WIKI
  - Problem: silent majority

History



Overview

- Introduction to GXL
- Emergence of GXL as a Standard
- Participating Communities
  - Software reverse engineering
  - Graph transformation
  - Graph drawing
- Lessons for Evolving Knowledge Artifacts

Reverse Engineering

- Strong precedent, established need
  - Tool sharing already occurring
  - Discussions about a standard format since 1996
- Commitment by core of researchers
  - Many visible and vocal, many silent
- Champions
  - Holt and Winter
  - Some from outside of GXL design team
- Feedback through meetings
  - Semi-annual meetings by Canadian researchers
  - Annual meetings at WCRE

Reverse Engineering (cont.)

- Community was ready
  - Ratified standard *before* GXL 1.0 was released

Graph Transformation

- Weak precedent, strong desire
  - Tool sharing was a goal
  - Wanted to apply graph transformation to other domains
- Commitment by small group of leading researchers
  - Small communities, others quickly followed
- Champions
  - Schürr, Taentzer, Varró, Winter
- Feedback through meetings
  - Short development cycle, many meetings over 2 years

Graph Transformation (cont.)

- Community was willing
  - GXL used as basis for GTXL
  - Ratified in March, 2001

Graph Drawing

- Very strong precedent, recognized need
  - Many, many existing formats
  - Creating a format is a rite of passage for researchers
- Interest strong, commitment weak
  - People were intrigued and were willing to be convinced
- Champions
  - Small number of champions, none of them leading or senior researchers
  - None of them part of GXL design team
- Feedback through meetings
  - Annual meetings, but many design decisions made without communication in between

Graph Drawing (cont.)

- Community is still looking for a format.
- Invited to join GXL effort, but lacked appropriate champions who had time
- Ideas were taken from successful graph formats.

Communities and Formats

- Claim: Standardization provides insight into the underlying technical problem and participating scientific communities.
- Synergy
- Ideas from each of the communities were absorbed into GXL.
  - Recall comments on standardization
- GXL was a “blank canvas question” for the three communities.
- Schema-based data exchange not adopted.

Dimensions of Software Data

Programming Languages
- single Languages (Ada, C, C++, Cobol, Java)
- multi-language systems

Level of Abstraction
- AST-level
- Architectural level

Relational Aspects
- Dataflow, Controlflow, ...
- Includes, Calls, Uses, ...

- Attributes
  - Inherits, color, location, …

reverse engineering
reengineering
program comprehension

Community Culture

- Membership
  - Number, participation level
- Leadership
  - Number, status, visibility
- Significant research problems and peer evaluation

Overview

- Introduction to GXL
- Emergence of GXL as a Standard
- Participating Communities
  - Software reverse engineering
  - Graph transformation
  - Graph drawing
- Lessons for Evolving Knowledge Artifacts

Summary

- Claim: A data format is a knowledge artifact.
- Claim: Successful standardization is a community-driven process.
- Claim: A successful standard data format is a community-driven knowledge artifact.
- Claim: Standardization provides insight into the underlying technical problem and participating scientific communities.