Linguistic and Architectural Requirements for Personalized Digital Libraries

Joachim W. Schmidt, Gerald Schröder, Claudia Niederée, and Florian Matthes

Technische Universität Hamburg-Harburg, Arbeitsbereich Sofwaresysteme, Harburger Schloßstraße 20, D-21079 Hamburg, Germany

Received December 10, 1996; accepted December 10, 1996

Abstract. Our vision of digital libraries is influenced by our experience with systems for persistent and networked object management and with polymorphic programming languages for their implementation. When viewed from this perspective, the essence of digital libraries can be captured by the following three essentials:

- the content of a digital library is represented by two kinds of information entities; on the basic level there are information tokens as supplied by information providers on the net; value is added to such tokens by individually constructing information artifacts over them with the goal of information consumer satisfaction;
- the services required for artifact construction and use – on the information level as well as on the level of the software artifacts required for these processes – rely heavily on powerful binding environments for multi-medial, persistent and networked information;
- the processes of artifact construction and use are in themselves valuable sources of information about artifacts; for the exploitation of such process information, digital libraries employ advanced tracing environments.

We derive linguistic and architectural requirements for digital libraries from these above essentials. On the language level we concentrate on generalized requirements for the typing, binding and scoping of library entities and services. On the system level we discuss architectural requirements in terms of orthogonal persistence, open extensibility, platform independence, mobility and reflection.

We present Tycoon [Matthes and Schmidt 1992; Matthes et al. 1995], a polymorphic, higher-order language and its system, and demonstrate its potential for digital libraries. We evaluate Tycoon’s rich conceptual basis (data, functions and threads), library-based extensibility, powerful binding mechanisms, its orthogonal persistence and its capability of network-wide data, code and thread migration.

We conclude by referring to an interdisciplinary digital library project in Art History Research based on icons, texts and data. Here, Tycoon effectively supports the process of individually customizing and scaling library services thus generalizing the notion of a query language into that of a persistent personal reference library.

1 Introduction

Digital libraries are coming into their own at a time when computer and network technology is improving rapidly resulting in increased global acceptance of digital information and communication services.

As a first consequence, an enormous quantity of information on virtually any medium is now being exchanged between a world-wide community of information providers and consumers. This potential will virtually revolutionize all information processing activities. Digital libraries are intended to provide the software which helps to structure this global information space and to improve its use.

World-wide dissemination of digital contents is, however, not the only achievement of the rapidly improving information and communication technology. Liberating computer scientists from too narrow technological restrictions also gives rise to the development of highly advanced models, languages and systems. If not, we are filling new wine into old barrels thereby creating another Pandora’s box of legacy problems for the upcoming information age.

Our vision of digital libraries is founded on both of the above achievements. We firmly believe that the great potential of the global information space can be exploited only if digital libraries make best use of highly advanced models, languages and systems. If not, we are filling new wine into old barrels thereby creating another Pandora’s box of legacy problems for the upcoming information age.

Based on our vision of digital libraries outlined in Section 2 we will present in Section 3 the essential linguistic and architectural requirements which we consider necessary for the construction and use of such systems. In Section 4 we will evaluate Tycoon [Matthes and Schmidt 1992; Matthes et al. 1995], the language and its system platform, and demonstrate its potential for digital libraries. Section 5 presents our initial experience with the Warburg Electronic Library, a cooperation project with the Art History community which involves data, texts and large amounts of iconic images.
In our view, the final purpose of a digital library is to improve the exploitation of the globally networked information universe with a clear focus on the particular information needs of an individual user and his task at hand [Van House 1995]. In other words, a digital library is the software needed to bridge the gap between the virtually infinite amount of information out there on the net and the very specific information needs of a particular person and task.

With this goal in mind, we distinguish on the level of library content two kinds of library entities:

*Information tokens* represent the information content which is collectively provided by the net; information tokens are multi-media and may represent texts, data, images, sound, videos etc. [McNab et al. 1996; Li et al. 1996]. In addition to information tokens digital libraries also contain software tokens, like applets or generic functions (such as filters, constructors, evaluators, presenters, binders or wrappers). *Information artifacts* are (recursively) created views over information tokens. Selecting, structuring, combining, and annotating are typical steps in this process. Artifact construction aims at adding value to information and thereby increasing the satisfaction of information consumers. Information artifacts may be image tokens associated with texts on their content as well as entire multi-media libraries including sound and videos.

Two main library processes can be identified: value-adding artifact construction and personalized exploitation. Although this paper concentrates on information artifacts it should be emphasized that the software required for personalized library processes puts heavy demands on software customization. Indeed, there is a striking similarity between processes which personalize the library’s information content and those which customize its software required for doing so. Such software is also represented by software tokens out of which customized software artifacts can be constructed which are finally bound and applied to information artifacts. In our closing remarks we will come back briefly to this similarity between information and software libraries.

These library processes are supported — on the information as well as on the software level — by two kinds of library environments:

*Binding environments* enable artifact construction and exploitation on the two interwined levels of information artifacts and information handling software artifacts. Binding environments provide a wide range of binding capabilities and support their safe and disciplined use in heterogeneous, open, networked and multi-media settings.

*Tracing environments* are highly desirable because library processes themselves are sources of valuable information on artifacts. Tracing such processes and extracting context information (on the who and when, the what and where, the why?), but also on software

2.1 Information Tokens

Information tokens represent the basic multi-media content of digital libraries and are provided by a digital and networked information universe. Their main purpose is to conceptualize information on an elementary level completely independent of any specific contexts in which information is referenced, aggregated and consumed.

Examples of information tokens are

- pictures in digital formats such as JPEG or GIF;
- texts in formats such as ASCII, Postscript or Word;
- time dependent tokens such as audio and video sequences in QuickTime or MPEG formats;
- software tokens such as JAVA applets or generic sorting functions in source or compiled bytecode or even native (machine-dependent) code.

To best serve their role as basic building blocks in value-adding digital libraries, tokens should be unbiased in their representation. Furthermore they should be self-describing, equipped at least with basic methods for presentation, copying, migration, storage etc. Note that the information tokens may inherently provide further more specific methods. A Word document, for example, may come with methods for reformatting or generating a table of content. For software tokens the basic methods may be compilation, execution or interpretation.

Independent of their content and the platform on which they reside, tokens have to share the following basic properties:

- digital representation
- atomic view, i.e., context-independent interpretation
- unconstrained content described by automorphic typing.

Automorphic typing allows the basic classification of information tokens, stresses their self-contained character of the information tokens and avoids restriction to predefined formats.

Information tokens have to be handled uniformly, independent of their content, lifetime or location in the network space. Therefore, the following additional properties are requested uniformly over all kinds of information tokens:

- network-wide, stable identity
- universal referentifiability via the net.

Any specific demands on information which are derived from the needs of individual users or user groups are captured by the concept of information artifacts.
Information artifacts are constructed (recursively) from tokens with the intent to add, step by step, value to the information on its way from the network of information providers to the individual information consumer. The structure of artifact constructing processes is shown in Figure 1. Information tokens provided by a networked information universe are composed into information artifacts, a process guided by the two related principles that value be added by information composition, but also by personalization [Röseheisen et al. 1995; Röseheisen et al. 1994]. Conceptually, a digital library has manifold artifacts visible to information consumers with high emphasis on personalized artifacts customized for the consumers preferences and tasks.

Typical examples of personalized information artifacts are individual collections of text and image tokens (references or copies) augmented by hyper-text artifacts and personal annotations. Artifacts may be attached to agents which search autonomously for further material on the net by employing text retrieval systems or SQL engines. Other examples are video sequences augmented by data on title, producer, costs, list of actors, comments and ratings etc.

Examples of software artifacts are application packages such as Oracle’s graphical query tools with GUI and SQL components or Microsoft’s application package Office composed of basic modules such as Word, Excel, and Powerpoint.

After all, effective exploitation of information artifacts relies on additional and more traditional library services [Graham 1995; Cousins et al. 1995; Paepcke 1996] such as

- information filtering capabilities (static and dynamic views)
- mobility in heterogeneous environments (platform independence)
- exchange between libraries (external gateways)
- persistent storability, multiple user support and recovery (database system support)
- security and accounting services (authorization and authentication).

However, it is the binding technology which, in our view, provides the core capabilities for artifact construction and use - on the information as well as on the software level. Therefore, binding requires particularly strong support in any modern digital library system.

2.3 Binding Environments

On the technical level, extended binding capabilities are most crucial for our approach to digital libraries. It is through binding that selected information entities can be organized into artifacts and that information consumers can access the artifacts of their choice. It is also bindings which connect information artifacts to software artifacts when value-adding software services are requested for artifact exploitation.

> have to be possible between all kinds of entities on an open set of media;
> must also be complete in the sense that all computational entities, data, code and threads can be freely combined by bindings;
> may be fixed over time or mutable on demand;
> may refer to transient or persistent entities;
> may involve local entities or reach out to remote sites of the net.

Bindings to entities internal to a system have to be defined and managed as well as external bindings to entities spread over heterogeneous systems. Artifact construction may require both kinds of semantics - copy semantics and reference semantics - via bindings; the transition between both has to be supported.

2.4 Tracing Environments

People, when working with libraries, usually follow certain patterns of use. Frequently they work top down, from vague ideas via keywords and catalogues to volumes and individual texts. Library users work continuously in the sense that they often resume sessions at the point of interrupt some time earlier. There also exist distinguished modes of operation which people apply when using a library:

- quick look-up of references to dictionaries or texts, or
- evaluation and memorizing of a concrete text performed in a series of sessions, or
- construction, maintenance and use of topic-oriented personal reference libraries over years.

The individual context in which such patterns are applied varies, of course, with the user and his task (the who and when, the what and where etc.). With digital libraries we are in a position to trace library processes on an individual level. We can discover individual patterns behind processes and provide environments to support them. Individual process execution contexts can be recorded and the information gained can be used for improving the library content as well as its usability [Böhme and Rakow 1994; Kashyap et al. 1996].

Typical information gained from process traces answers questions such as

- who is involved in artifact construction? when is it constructed?
- which digital libraries are consulted to find information? where do they reside?
- which queries are posted and what are the results of these queries?
- what alternatives or versions exist for artifact construction?
which design decisions lead to the artifact?

Some of this information (see Figure 2) can only be given explicitly by the person who constructs the artifact (e.g., information on the why?). Other information can be deduced automatically by the tracing process attributes (e.g., owner, time), statement sequences (e.g., retrieval operations) or process bindings (retrieval results).

We conclude our vision of digital libraries by contrasting them with traditional database systems: databases contain homogeneous data (e.g., relations) and database architectures provide a centralized view on data and software (global schema; distribution transparency). Furthermore, databases are accessed by isolated transactions (ACID principle). Digital libraries differ substantially from databases in all three dimensions just mentioned: they access heterogeneous information represented by multiple models on multiple media; their architecture is a dynamic and highly personalized construction of information and services, and their users work with contexted processes which are traced and evaluated by the library environment. Some of these extended requirements of digital libraries are addressed in the research area of information integration and heterogeneous databases (see e.g., [García-Molina et al. 1995; Hammer and McLeod 1993]).

In the following section we discuss some of the linguistic and architectural demands on software platforms on which digital libraries are best developed, maintained and used.

3 Software Requirements for Digital Libraries

Our view of digital libraries imposes several requirements on software platforms on which such libraries are developed, used and maintained. Major demands result from the following properties in digital library systems:

- safe extensibility to an open set of new media types, novel library services, extended user requirements etc. [Nürnberg et al. 1995]
- scalability from private stand-alone libraries to shared group libraries or globally networked libraries
- global connectivity with free data exchange, open systems communication, unrestricted software migration etc.
- customizability for personalized, value-added information and software artifacts.

Some of these demands involve language requirements, others imply architectural requirements for the digital library system itself. Both sets of requirements are clearly not completely independent of each other as we will see with requirements such as binding or reflection.

3.1 Linguistic Requirements

Our vision of digital libraries raises specific demands on the naming, typing, binding and scoping capabilities of the language of our choice. We argue here for some fairly novel language concepts by which data, code and threads become first-class citizens, i.e., players with equal linguistic rights.

3.1.1 Typing

Digital libraries essentially impose three important demands on the underlying typing system:

- type expressiveness for extensive model definition,
- type control for safe model execution, and
- type inspection for generic programming.

Types for modeling. The information modeling needs of a digital library are abundant and may vary over its
lifetime. Consequently, languages with a fixed, built-in set of models do not suffice. Instead, an add-on approach to model definition is requested, a demand which can best be met if such models can be expressed by the language itself. Such a demand leads directly to languages with rich polymorphic type systems, (recursive, parametric) type constructors, subtyping and higher-order functions.

Tasks to be addressed by model extensions are bulk data management [Schmidt 1977; Ackerman and Fieldsing 1995], multi-media modeling, safe service integration, etc. There is a particular need for a wide range of bulk or collection types such as lists, sets, keyed sets and bags. The framework should provide an add-on bulk type library which satisfies different requirements for search, insert, order or store. Bulk types dynamically added on should, at the same time, be safe, disallowing, for example, the insertion of image tokens into book collections.

In our approach expressive typing schemes come into effect on the artifact level; tokens are automorphically typed (i.e. self describing) bit vectors, a representation which simplifies generic low-level operations such as copying, storage, migration or garbage collection. Furthermore, the automorphic typing enables more specific methods, e.g., for token presentation and basic manipulations.

**Types for control.** Types capture constraints on computations. Good type systems allow a rich set of constraints to be expressed and to be checked as early as possible. Type checking is usually done statically, i.e., at compile-time: at that time the price for type checking is cheapest and type violations can most easily be fixed. However, long-lived and distributed systems, such as digital libraries, are frequently faced with constraints which arise at different sites or change during their lifetime. This requires more dynamic approaches to typing. If, for example, values are transmitted (via files or communication channels) between independently developed digital libraries, there is no common scope in which a static type check could be performed to guarantee compatibility between data and programs. For tasks like these, (value, type) pairs have to be available at run-time, and a rich functionality is required for run-time evaluation and exception handling.

**Types for genericity.** Another need for dynamic types results from the desire to implement generic functions with type-dependent behavior. Such functions take a type representation, usually along with a value of this type, inspect the type and exhibit different behavior depending on the type. Type inspection allows, for example, iteration over the attribute types and attribute values of an aggregate or construction of an aggregate from a list of typed bindings. Such possibilities are required, for example, for generic library browsers capable of displaying and manipulating artifacts of any type.

### 3.1.2 Binding capabilities

Different kinds of bindings between information artifacts and tokens as well as between information and software artifacts are required. It should be possible to combine them orthogonally:

**Static and dynamic.** In traditional applications bindings to entities are stable over time. For digital libraries such static bindings are, however, not sufficient since artifacts are created dynamically and may be updated. Dynamic bindings are functions (queries or other computations) that are dynamically evaluated and establish bindings on demand.

**Internal and external.** Internal bindings refer to entities which are under the control of the active system while external bindings refer to entities outside. External bindings frequently require special treatment because essential properties of the active system may not apply to external entities. Typical examples of such properties are orthogonal persistence or transactional access.

**Local and remote.** Local bindings refer to entities on the site where the process is currently executing. Digital libraries also require remote bindings which can refer to entities that reside on a different site in the network. Remote bindings imply automatic transfer to the local site if the bound entity is referenced, e.g., for display.
Digital libraries have to be prepared to accept services from ‘alien’ systems. The diversity of services required by a digital library could never be provided without external support. The demand for novel services is driven by both sides, providers and consumers, and is expected to grow rapidly with the number of libraries on the net. External services may be required, for example, in the following situations:

- external tokens – images, sounds, but also native code – can frequently only be interpreted by software which is only available on the site on which the token resides;
- some sites specialize in services not available at the home site; examples are expensive printing or scanning facilities;
- many standard services, e.g., for presentation, compression, encryption, communication, etc., are factored out to generic subsystems; it is often the case that the complexity of a service demands a specialized service provider; examples are information retrieval and SQL engines;
- finally, legacy awareness requires that external service integration is approached on a general level.

On the modeling level safe service integration can be achieved by expressive typing (see previous subsection); on the system level it requires gateways and call/callback
interfaces to the external service’s languages (C, C++, ...). In combination, both requirements support the definition and implementation of new software libraries which integrate external services safely into our language of choice. In addition to what is required for service integration in general, standards for export interfaces and for exported references to library entities are necessary. Finally, each digital library also should be prepared to export its services to others and to act as information provider in the net.

3.2.3 Platform Independence

Digital libraries require that information and services must be made available network-wide on heterogeneous platforms (see Figure 3). Library exploitation can be improved considerably if a library’s core functionality for searching, querying and storage is provided platform-independently. This means that data, code and thread representations abstract from all platform-dependent details. Technically speaking, this leads to simplified abstract machine architectures based on bytecode representation and interpretation as well as to run-time systems with improved portability, minimal in size and composed of standardized components.

3.2.4 Migration Technology

The usability and autonomy of digital libraries is substantially improved by agents [Birmingham 1995] capable of migrating freely in open networks to visit data sets and collect information [Balabanovic and Shoham 1995].

3.2.5 Reflection

There are numerous reasons why digital artifacts may have to be occasionally rearranged when used in networked and long-lived environments. Software migration, integration, (re-)binding, tracing are some such reasons, but also wrapping, persistent storage or optimization. Frequently, such tasks are highly regular and can be systematically performed by algorithms.

Reflection is a powerful yet (relatively) safe technology for the algorithmic inspection and readjustment of digital representation. Reflective program analysis and transformation has to rely on detailed knowledge about the programming language model and its implementation. Reflection may not only be used for generating or altering data or code (at compile-time or run-time) but also for transferring compile-time information (e.g. on types) to run-time. Tasks for which reflection is required in our setting are

> binding environments: reflective manipulation of bindings may, for example, transform local references to remote ones or replace remote references with copies;

> tracing environments: reflection can exploit information in the environment in which threads are executed (e.g. date, user-id, software version numbers, etc.) and attach such meta-information to the executing thread. Such information is collected automatically as a side-effect of user actions. This information collection process can be parameterized by user-defined filters according to user preferences.

4 Tycoon as a Framework for Digital Libraries

The goal of the Tycoon project [Tycoon 1992] is to provide modeling flexibility as well as system stability for multi-functional, long-lived application systems operating in heterogeneous, open and networked environments. Tycoon contributes to this goal on two levels [Matthes et al. 1995].

> The Tycoon language is a persistent polymorphic programming language with an elaborate higher-order type system [Matthes and Schmidt 1991; Schmidt et al. 1993; Schmidt and Matthes 1993; Matthes and Schmidt 1993].

Tycoon: Typed communicating objects in open environments.
Let BookComponent = Tuple
case text with content : String
case image with content : jpeg

Let Book = Tuple
authors : set.T(Author) title : String

The Let construct introduces type bindings. In the example above, the two variants (text and image) of the tuple type Book contain a field labeled content which is of type String in the case of text and of type jpeg in the case of image.

The second type defines a tuple with three fields (authors, title, contents). The field authors is defined as a set of authors, where set.T is a polymorphic type operator that takes the element type as actual parameter.

By structural subtyping a (recursive) type Publication becomes a subtype of Book so that all operations on books, e.g. the displaying of the contents, are also applicable to publications:

Let Rec Publication <: Ok = Tuple
authors : set.T(String) title : String
contents : list.T(BookComponent)
references : list.T(Publication) end
let displayBook(book : Book) : Ok = ...
let napoleonicWars : Publication = ...
displayBook(napoleonicWars)

The notation <: Ok in the type definition defines a super type needed for type checking of recursive types, in this case the super type of all unparametrized types, Ok. The let construct introduces value bindings. displayBook is a function taking one parameter of type Book and returning nothing (represented by the type Ok). napoleonicWars is a value of type Publication.

Generic structures such as sets or lists can be defined within TL through its higher-order polymorphic type system [Schmidt and Matthes 1994; Matthes and Schmidt 1991]. In the following example the recursive type operator for lists depends on an element type E so that lists of type T(E) accept only elements of type E:

Let Rec T(E <: Ok) <: Ok =
Tuple case nil case cons with head : E tail : T(E) end
let new(E <: Ok) : T(E) = ...
let cons(E <: Ok element : E list : T(E)) : T(E) = ...

The functions new and cons take a type parameter E constraining the types of the actual values passed to the functions resp. that describes the result value returned by the functions. This enables static type checking of generic operators.

TL supports dynamic types and dynamically-typed values. A dynamic value is a pair of a value v and a run-time representation t of its type. If a dynamic value component v is extracted, its associated type representation t can be inspected. This enables a boolean subtype test whether t is a subtype of a given supertype T. Such functionality is required to assure, for example, that an artifact transferred from an external site is a book:

(* -- transfer a dynamic value *)
let dynamicValue : dynamic.T = receive(channel)
(* -- test if this value is a book *)
let book : Book = dynamic.be(Book dynamicValue)
(* -- book is now statically typed *)
displayBook(book)
4.1.2 Binding Capabilities

Tycoon supports a wide variety of binding concepts covering a substantial range of what is requested in Section 3.1. The more traditional binding alternatives include immutable, mutable, static and dynamic bindings to types and values:

(* -- type binding: *)
let Book = Tuple... end

(* -- value binding: *)
let napoleonicWars = tuple... end

(* -- mutable binding: *)
let var actualBook = tuple... end

(* -- static binding: *)
let bookMark = actualBook

(* -- assignment: *)
actualBook := tuple... end

(* -- dynamic binding: *)
let displayBook(book: Book) : Ok = ...

displayBook(napoleonicWars)

In Tycoon it is completely transparent whether a binding refers to a local or a remote entity [Mathiske et al. 1995a; Mathiske et al. 1995b]. If an entity leaves the local scope in which it is bound, two alternatives exist: Either the entity (and its closure) is copied or the binding is replaced by a remote reference. On request remote references may be eliminated by copying.

Bindings to external functions are established by the predefined bind function. Calls of external functions are indistinguishable from calls of internal functions. Tycoon takes care of the machine-dependent part of the parameter conversion and call mechanisms. The following example shows the binding to a machine-dependent function which displays JPEG images:

let displayJPEG = bind(Fun(JPEG)) : Ok

"imageLib" "display_image"

displayJPEG(image)

The type Fun(JPEG) : Ok describes the signature of the external function. This is, again, a type parameter that enables static type checking.

It should be noted that call and binding transparency is a language requirement with heavy demands on the architecture of the system.

Artifacts contain data but may also require code and thread components. Code entities are used to construct artifacts dynamically ‘just in time’. Each time the subsequent function newsOnElection is evaluated the Reuters news service is asked for the latest news on the election:

let newsOnElection() : set T(News) = begin
  let reuters = openConnectionTo("Reuters News Service")
  searchFor(reuters "Election")
end

let actualNews = newsOnElection()

Even more value may be added to artifacts by higher-order functions expressing specific personal preferences, such as an interest in news which mention both candidates:

let newsOnElection(pref(News): Bool) : T(News) = ...
let bothCandidates(news: News): Bool = news.includes("Clinton") and news.includes("Dole")

let actualNews = newsOnElection(bothCandidates)

With threads components artifacts can work offline based on local state information. To collect only the latest news from a news service a thread may be employed which remembers the time it last collected the news:

let collectLatestNews(): Ok = begin

let var lastCollectionTime = never
let isActualNews(news: News): Bool = news.time > lastCollectionTime

loop

  let latestNews = newsOnElection(isActualNews)
  lastCollectionTime:= now()
  deliverNews(news)
  thread.sleep()

end

let newsThread = thread.start(collectLatestNews)

Our newsThread collects (in parallel to other activities) all news available since its last wakeup; then it goes to sleep until reactivated by thread.wakeup(newsThread).

Threads are first-class enitities in Tycoon and can be observed (traced) reflectively. In this way context data of artifact construction processes can be created and stored persistently [Mathes and Schmidt 1994]. With Tycoon’s reflective capabilities such processes can be parameterized by user-prefered filters for the information to be attached to the artifact. This may not only add value to the constructed artifact but also help improving the artifact construction process.

4.1.4 Exception Handling

Since unexpected events are frequent in open systems, Tycoon provides strong support for the handling of exceptional situations. If an exception is raised, execution stops and continues where the last exception handler was
The architectural requirements for digital libraries can be combined with modules for distribution, versioning, etc. Tycoon on System Architecture

Tycoon applications, frequently, are very large. As a result, they can be structured into interfaces, modules and libraries which restrict the scope of bindings [Schmidt et al. 1995]. Tycoon libraries are nested high-level entities that declare a definition order for interfaces and modules so that they can only use identifiers which were defined previously, preventing cyclic dependencies. Interfaces export public identifiers while modules define the bindings for these identifiers using also private bindings. Through nesting and hiding mechanisms the scope of identifiers can be controlled.

Currently, an orthogonal suite of scoping primitives is under investigation which specializes in the requirements of personalized scopes in open networked environments. It is provided by a generic Tycoon module that may be combined with modules for distribution, versioning, access control [Rudloff et al. 1995], etc. thus enabling the flexibility required by personalized digital libraries.

4.2 The Tycoon System Architecture

The architectural requirements for digital libraries can be seen from two perspectives.

- Since digital libraries hold a vast variety of information content and provide extended information handling services there is an intrinsic complexity involved in libraries seen as a performing software system.
- Complexity also results from the fact that library systems will never stay stable; instead it has to be anticipated that libraries will develop during their virtually infinite lifetime, even more so when connected to a virtually unlimited network space.

Tycoon reduces the complexity of both the above tasks, first by enabling the construction of systems with a higher degree of regularity (through orthogonal persistence, platform independence, migration technology etc.), second by supporting system evolution and change management (open service integration, scalability, reflection). The layered architecture of the Tycoon System (see Figure 4) scales from single-user PC-based applications to distributed applications in open heterogeneous networks.

4.1.5 Scopes

Tycoon applications, frequently, are very large. As a result, they can be structured into interfaces, modules and libraries which restrict the scope of bindings [Schmidt et al. 1995]. Tycoon libraries are nested high-level entities that declare a definition order for interfaces and modules so that they can only use identifiers which were defined previously, preventing cyclic dependencies. Interfaces export public identifiers while modules define the bindings for these identifiers using also private bindings. Through nesting and hiding mechanisms the scope of identifiers can be controlled.

Currently, an orthogonal suite of scoping primitives is under investigation which specializes in the requirements of personalized scopes in open networked environments. It is provided by a generic Tycoon module that may be combined with modules for distribution, versioning, access control [Rudloff et al. 1995], etc. thus enabling the flexibility required by personalized digital libraries.

4.2 The Tycoon System Architecture

The architectural requirements for digital libraries can be seen from two perspectives.

- Since digital libraries hold a vast variety of information content and provide extended information handling services there is an intrinsic complexity involved in libraries seen as a performing software system.
- Complexity also results from the fact that library systems will never stay stable; instead it has to be anticipated that libraries will develop during their virtually infinite lifetime, even more so when connected to a virtually unlimited network space.

Tycoon reduces the complexity of both the above tasks, first by enabling the construction of systems with a higher degree of regularity (through orthogonal persistence, platform independence, migration technology etc.), second by supporting system evolution and change management (open service integration, scalability, reflection). The layered architecture of the Tycoon System (see Figure 4) scales from single-user PC-based applications to distributed applications in open heterogeneous networks.

4.2.1 Persistence

Tycoon provides orthogonal persistence for data, code and threads. The Tycoon Store Protocol, TSP [Matthes et al. 1996], defines a uniform, data model-independent call-level interface to multiple (commercially distributed) persistent object stores abstracting from implementation details of the underlying persistent stores. A typical TSP server is implemented by a store adaptor which maps TSP data structures and functions to data structures and operations of an existing object store.

TSP contributes significantly to Tycoon's system scalability since TSP clients can decide between different

- garbage collection strategies
- object faulting mechanisms
- error recovery, logging and persistent savepoint mechanisms
- commit protocols (for distributed systems)
- security and authentication support

or choose between a single- or multi-user environment.

A primary design goal of TSP is to provide efficient data storage independent of the data and language model used by TSP store clients. TSP supports polymorphically typed models where store objects may contain values of different types. TSP therefore uses an untyped
4.2.2 Open Service Integration

External services, e.g., visualization systems or database systems, can be integrated into the Tycoon system through bindings to external functions [Schmidt and Matthes 1993]. The binding mechanism allows the Tycoon system to call external functions as well as it allows the external system to call back the Tycoon system via higher-order functions. Examples are event-controlled windows systems that call application-defined functions when a button is pressed or a window closed etc. A service can be implemented as a simple thread that is pressed when a call back is needed.

Such services are wrapped by Tycoon in portable and type-safe functions exploiting the full polymorphic power of the Tycoon type system even in cases in which external functions are completely untyped. When calling external functions from Tycoon or, in reverse, calling (back) Tycoon functions from external systems, parameter conversion is handled by the Tycoon run-time system. The type system of Tycoon is designed to ensure that all necessary type conversions are performed correctly.

The following example shows an interface `SQL` that exports polymorphic functions to access different SQL database systems:

```haskell
interface SQL import ... export
    error : Exception with sqlError : String end
    Table (E, K < , Tuple end) < , Ok
    ... openTable (Dyn E, K < , Ok
        tableName : String) : Table (E, K)
    ... lookup (E, K < , Ok
        from : Table (E, K) key : K) : E
    ... selectFromWhere (E, K, R < , Ok
        project (E) : R
        from : Table (E, K) where (E) : Bool) : Iter.T (R)
    ...
end
```

Tycoon libraries hold, for example, two different modules `ingesSQL` and `oracleSQL` which implement the above interface by bindings to the dynamic SQL call interfaces of the Ingres and Oracle relational database management systems. The `selectFromWhere` function, for example, is used to retrieve a row from a table with a given key. It can be used in a `from` clause to select rows from a table that match a given condition. The `error` exception is an `Exception` with a message.

The following example shows an interface `SQL` that exports polymorphic functions to access different SQL database systems:

```haskell
interface SQL import ... export
    error : Exception with sqlError : String end
    Table (E, K < , Tuple end) < , Ok
    ... openTable (Dyn E, K < , Ok
        tableName : String) : Table (E, K)
    ... lookup (E, K < , Ok
        from : Table (E, K) key : K) : E
    ... selectFromWhere (E, K, R < , Ok
        project (E) : R
        from : Table (E, K) where (E) : Bool) : Iter.T (R)
    ...
end
```

Tycoon libraries hold, for example, two different modules `ingesSQL` and `oracleSQL` which implement the above interface by bindings to the dynamic SQL call interfaces of the Ingres and Oracle relational database management systems. The `selectFromWhere` function, for example, is used to retrieve a row from a table with a given key. It can be used in a `from` clause to select rows from a table that match a given condition. The `error` exception is an `Exception` with a message.

4.2.3 Platform Independence

The Tycoon Virtual Machine (TVM) is an abstract call interface above the TSP layer that defines a bytecoded instruction set based on a higher-order, functional execution model. TVM bytecode is either interpreted by a virtual machine or is compiled on the fly into target machine code. The TVM interpreter and its associated run-time system are written in ANSI-C.

The platform-independence of the TVM model makes it possible to dynamically transfer portable bytecode between heterogeneous nodes in distributed digital libraries without recompilation. Utilizing TSP's linear external data representation (TXR), it is also possible to migrate a thread across system boundaries [Mathiske et al. 1995a; Mathiske et al. 1995b; Matthes and Schmidt 1994].

For example, the higher-order query function `newOnElection` can be rewritten in the following way for
The Tycoon system supports different manners of reflection. By linguistic reflection [Stemple et al. 1991; Stemple et al. 1992] the execution of the compiler (compile-time reflection) or the application (run-time reflection) can be influenced.

Compile-time reflection is achieved by executing user-defined code during compilation. Dynamic types, for example, are implemented by compile-time reflection. The compiler collects type information and stores it persistently for run-time use. Similarly, operations on compile-time information are performed as, for example, by the Repeat construct which repeats a list of Tycoon signatures (for the definition of Book, see above):

```plaintext
let createBook(Repeat Book) : Book = ...
```

Here, Repeat Book is reflectively replaced by

```plaintext
authors : set.T(Author) title : String ...
```

4.2.4 Migration Technology

Even more flexibility is achieved in Tycoon because threads can migrate between the nodes in a network [Mathiske et al. 1995a; Mathiske et al. 1996]. This enables, among others, the implementation of agents. An agent may include the trace of the search in its search process, e.g. the items it has already found, the number of nodes visited or the nodes it has been recommended to visit:

```plaintext
let sitesToVisit = set.new(:Site) let homeSite = agent.thisSite() let sitesVisited = set.new(:Site) let local : SiteData = agent.migrateTo(nextSite) set.insert(sitesVisited nextSite) set.include(newsFound local news.search("Election")) let newSites = set.exclude(local sitesRecommended sitesVisited) set.include(sitesToVisit newSites) end
```

```plaintext
agent.migrate(homeSite) show(newsFound)
```

4.2.5 Reflection

The Tycoon system supports different manners of reflection. By linguistic reflection [Stemple et al. 1991; Stemple et al. 1992] the execution of the compiler (compile-time reflection) or the application (run-time reflection) can be influenced.

Compile-time reflection is achieved by executing user-defined code during compilation. Dynamic types, for example, are implemented by compile-time reflection. The compiler collects type information and stores it persistently for run-time use. Similarly, operations on compile-time information are performed as, for example, by the Repeat construct which repeats a list of Tycoon signatures (for the definition of Book, see above):

```plaintext
let createBook(Repeat Book) : Book = ...
```

Here, Repeat Book is reflectively replaced by

```plaintext
authors : set.T(Author) title : String ...
```

In summary, Tycoon is a powerful framework which provides much of the functionality required to make multi-functional and multi-media application systems persist in turbulent environments such as open, heterogeneous and networked information infrastructures.

5 The Warburg Electronic Library Project

The Warburg Electronic Library Project (WEL) began within the framework of an interdisciplinary cooperation [Niederée et al. 1996] between our group and the Art History department at the University of Hamburg. The goal of this cooperation is the examination, development and application of digital libraries for art history research purposes.

The development of this particular digital library focuses on two topics:

- understanding the special requirements of the application domain and their adequate realization;
- personalization of a working environment for art history research by tailoring personal digital reference libraries to the needs of individual information consumers and specific tasks.

5.1 The Application Domain

The application domain of art history is dominated by image material augmented by texts and multi-media artifacts such as films and audio documents. Art history
Linguistic and Architectural Requirements for Personalized Digital Libraries /1/3 research stresses the examination of themes, the identification of represented icons (symbols), events and persons and their classification according to these issues. In the handling of image material automatic image processing plays only a subordinate role. Most of the information has to be added by people in a value adding process.

This style of image handling is not restricted to art history. Other areas such as marketing and press archives are also primarily interested in themes and messages transported by their artifacts as well as the represented objects and persons.

The PI-Index

Political Iconography (PI) is the area of art history which examines political messages conveyed in images showing regents, politicians, ceremonies, political acts, etc. The underlying assumption of the PI is that the effects of political actions are not restricted to contracts and political documents but are also depicted in paintings, monuments and buildings.

The art history department has developed an elaborated ontology for the classification of images according to their political messages called the PI-Index. This ontology consists of a hierarchy of terms referring to politics, political acts, and social phenomena. It includes terms as varying as science, marriage, democracy, shepherd, and revolution.

The classification of image material according to this ontology cannot be done automatically. It is itself a result of scientific work in the area of art history. About 250,000 cards with photographs of paintings, etc. showing politicians, political acts and ceremonies, battles and social events from all epochs and countries are already classified according to this scheme. The classification is not disjoint, many cards are assigned to several terms of the index.

5.2 Tokens and Artifacts in the Warburg Electronic Library

Photographs or prints of paintings showing regents or related matters, associated texts, speeches of politicians, and documentary films about regents like Napoleon, form the relevant information tokens in this digital library. The tokens are created and included into the information universe through scanning or an equivalent digitalization process. A further source of tokens is text editing. Scanned image tokens are stored as bit vectors in formats like JPEG or GIF. In Tycoon types are associated to these bit vectors and routines, preventing the erroneous use of bit vectors and routines. The following example shows the association of a type T and the routines scan and display for JPEG images:

```haskell
let jpeg = tuple
    Let T = ...
    let scan = bind(Fun(T :T) :Ok "imageLib"
        "scan_jpeg_image")
    let display = bind(Fun(T :Ok "imageLib"
      "display_jpeg_image")
end
let jpegImage = jpeg.scan()
display.jpegImage
```

As mentioned in Section 4 the bind function in Tycoon is used to include external, machine-dependent routines scan_jpeg_image and display_jpeg_image that work on JPEG files.

On this level the tracing environment may provide information about the date of digitalization and the identification of the person creating this token.

The tokens are taken as starting points for the artifact construction. The image token showing Napoleon crossing the Alps, for example, is taken as a basis for an artifact as illustrated in Figure 5. The artifact contains a reference to the scanned painting (image token) and a thumbnail copy of it, as well as information about the artist, the title, and the date of the painting as well as the location where the original painting can be found and the source of the scanned print of the painting.

In addition, the artifact is accompanied by content-descriptive metadata: It is classified according to the ontology of the PI-Index and the regent pictured by the token, here Napoleon, is specified. This additional information can be considered as information tokens from a private information source.

The type representation of this artifact in Tycoon looks as follows:

```haskell
let PIArtifact = Tuple
    originalImage :Reference(jpeg.T)
    thumbnailImage :jpeg.T
    title :String
    date :Date
    ...
end
```

![Fig. 5. Napoleon crossing the Alpens at St. Bernhard](image.png)
The Warburg Electronic Library maintains a large collection of these artifacts together with services for their persistent storage, transmission, retrieval, and presentation. On the language level retrieved artifacts may be included into a list of type list.T(PIArtifact) and the associated thumbnail images may be displayed by using iteration abstraction provided as generic services by a Tycoon library.

list.forEach(artifactSelection fun(a : PIArtifact)
jpeg.display(a.thumbnailImage))

The library is augmented by an Oracle database containing historical data on regents of all centuries. The Tycoon SQL-gateway to Oracle databases is described in Section 4. Looking up historical data about Napoleon is accomplished by the following function of the SQL interface.

oracleSQL.lookup(regents "Napoleon")

In addition to image artifacts and structured data the Warburg Electronic Library also contains text collections, e.g., about political iconography and political events. These collections can be searched by an Inquery information retrieval engine. Employing the Tycoon Inquery gateway the collection can be searched in the same language framework as the Oracle database.

Inquery.eval(politicalEvents "Anything about Napoleon and Waterloo")

The query results thus can be easily combined, e.g., in order to prepare a presentation about Napoleon at Waterloo with historical dates from the database regents and text information from the text collection politicalEvents.

5.3 Personal Reference Libraries

Digital libraries provide multi-media material for groups of information consumers. To secure the success of a complex task a working environment tailored to the individual requirements of the consumer and to the specific task is necessary. The development of such personalized, cooperative working environments called personal reference libraries is currently being examined in our group. Personal digital reference libraries can be stored persistently and exchanged with other persons who work on the same or a similar topic.

An example of the construction of a personal reference library is the preparation of a publication and/or demonstration on equestrian portraits of regents throughout the centuries.

The ontology PI-Index includes the terms Reiterbild, i.e., equestrian portrait, and Herrscher, i.e., regent. The set of pictures and other material classified into both categories can thus be easily accessed using the services of the digital library.

Central to our view of digital libraries over networked information tokens is the value-added construction of information artifacts and their use in personalized information environments. The library processes of artifact construction and use are enabled and supported by advanced binding and tracing environments which place substantial demands on the linguistic and architectural

```java
list.forEach(artifactSelection fun(a : PIArtifact) jpeg.display(a.thumbnailImage))

The library is augmented by an Oracle database containing historical data on regents of all centuries. The Tycoon SQL-gateway to Oracle databases is described in Section 4. Looking up historical data about Napoleon is accomplished by the following function of the SQL interface.

oracleSQL.lookup(regents "Napoleon")

In addition to image artifacts and structured data the Warburg Electronic Library also contains text collections, e.g., about political iconography and political events. These collections can be searched by an Inquery information retrieval engine. Employing the Tycoon Inquery gateway the collection can be searched in the same language framework as the Oracle database.

Inquery.eval(politicalEvents "Anything about Napoleon and Waterloo")

The query results thus can be easily combined, e.g., in order to prepare a presentation about Napoleon at Waterloo with historical dates from the database regents and text information from the text collection politicalEvents.

5.3 Personal Reference Libraries

Digital libraries provide multi-media material for groups of information consumers. To secure the success of a complex task a working environment tailored to the individual requirements of the consumer and to the specific task is necessary. The development of such personalized, cooperative working environments called personal reference libraries is currently being examined in our group. Personal digital reference libraries can be stored persistently and exchanged with other persons who work on the same or a similar topic.

An example of the construction of a personal reference library is the preparation of a publication and/or demonstration on equestrian portraits of regents throughout the centuries.

The ontology PI-Index includes the terms Reiterbild, i.e., equestrian portrait, and Herrscher, i.e., regent. The set of pictures and other material classified into both categories can thus be easily accessed using the services of the digital library.

Central to our view of digital libraries over networked information tokens is the value-added construction of information artifacts and their use in personalized information environments. The library processes of artifact construction and use are enabled and supported by advanced binding and tracing environments which place substantial demands on the linguistic and architectural

```
Linguistic and Architectural Requirements for Personalized Digital Libraries

On various occasions in this paper we mentioned already the similarity between digital libraries with information artifacts and software libraries (e.g. [Meyer 1990; Meyer 1984]) with generic software tokens and customized software artifacts constructed for specific use. Based on our initial experience we expect that much of the functionality of the binding and tracing environment for information artifacts can also be used for the construction and use of the software artifacts required by digital libraries. The first-class status of data, code and threads in Tycoon provides a promising framework for such an approach.

On the technology level, we have a vested interest in the extension of our binding technology, e.g., by the concept of versioned bindings and by extended operations on structured collections of bindings.

The most relevant input, however, for the future development of digital libraries we expect from our cooperation with potential user communities which have both the relevant library contents as well as the pragmatics of how to use them. From their feedback we expect to gain the domain knowledge which is required to design and realize the adequate abstractions for future digital libraries.

Acknowledgement. This research was supported by Esprit Projects ISC-CAN-080 and 22552-PASTEL.


