

ForeCite: Towards a reader-centric scholarly digital library*

Thuy Dung Nguyen Min-Yen Kan Dinh-Trung Dang¹
Markus Hånse² Ching Hoi Andy Hong³ Minh-Thang Luong
Jesse Prabawa Gozali Kazunari Sugiyama Yee Fan Tan

Department of Computer Science, School of Computing
National University of Singapore

{nguyen14,kanmy,luongmin,jprabawa,sugiyama,tanyeeefa}@comp.nus.edu.sg
¹dangdinhtrung@gmail.com ²mhaense@gmail.com ³weelillad@gmail.com

ABSTRACT

We present ForeCite (FC), a prototype reader-centric digital library that supports the scholar in using scholarly documents. FC integrates three user interfaces: a bibliometric component, a document reader and annotation system, and a bibliographic management application.

1. INTRODUCTION

Scholarly digital library (DL) efforts to date have largely concerned themselves with the stewardship and collection of scholarly documents. DLs provide analytics for judging the quality of individual or collective research, including citation data, circulation and download counts, and information about individual authors, publication venues, grants, etc. In this sense, current DLs treat documents as opaque objects, and do not offer tools to aid their reading and utilization. However it is clear that such tools are needed. Annotation frameworks, bibliographic managers, notetaking software and onscreen readers are all existing technologies that support reader-centric functions, but are all independently implemented, creating a piecemeal solution. We present ForeCite (FC), a platform that supports a cross-section of these common functions.

2. FORECITE SYSTEM ARCHITECTURE

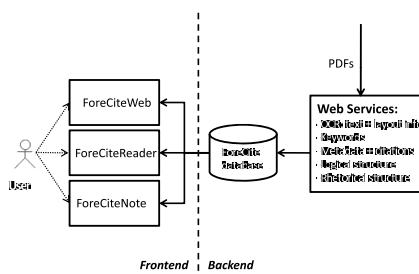


Figure 1: ForeCite architecture.

Figure 1 gives a high level overview of FC’s architecture. FC backend components are structured as web services to

*This work was partially supported by a National Research Foundation grant “Interactive Media Search” (grant #R-252-000-325-279).

allow a more modular and flexible design. Documents first undergo OCR to extract both text and layout information. Other web services then operate on the extracted data to provide bibliographic metadata, keyphrase extraction, reference string, logical structure and rhetorical structure parsing. These extracted data are stored in the FC database, running MySQL. This database forms the core of the FC system, providing the necessary data for the frontends. An FC API allows third party applications to provide new annotations, so that other services can easily extend FC.

In FC we have implemented three web-based user interfaces to help with the readers’ primary activities: exploration, reading, and synthesis.

FCWeb. An interface that assists users in finding and exploring scholarly documents, similar to CiteSeerX and ACM Portal. For each document, FCWeb provides bibliographic information such as citation and reference information. FCWeb also shows user annotations and reviews made in the other two components.

FCReader. A reading environment that supports collaborative annotation and content analytics, similar to Google Books and Adobe Acrobat. FCReader allows users to annotate, highlight and view other’s collaborative annotations. What sets FCReader aside from reading utilities is its integration of analytics. FCReader can show different annotation layers that correspond to automated document analysis: automated keyphrase extraction, logical as well as rhetorical structure. These functionalities help readers quickly digest and locate content within documents themselves.

FCNote. A bibliographic manager for notetaking and organizing personal scholarly document collections. FCNote is presented as a self-contained HTML file with Javascript which enables the notetaking functionality and is usable in any major web browser, either on- (web connected) or off-line (on the local file system). In complement to FCReader, FCNote focuses on a user’s annotations of a paper and allows users to see annotations across multiple papers. FCNote also offers the local management of PDF documents that are stored locally on users’ PCs.

A core consideration of FC is to ensure that three interfaces are supportive of each other and integrated, in the sense that primary information in one interface can serve and be shown as secondary information in another. Public annotations by users appear in all three interfaces. Reference metadata for a paper appears in both FCWeb as well as FCNote. As ongoing work, we continue to strengthen the ties between the interfaces, to better support secondary tasks without leaving the user’s primary interface.