Applying Semantic Web Services to Bioinformatics: Experiences Gained, Lessons Learnt

Phillip Lord[1], Sean Bechhofer[1], Mark D. Wilkinson[2], Gary Schiltz[3], Damian Gessler[3], Duncan Hull[1], Carole Goble[1], and Lincoln Stein[4]

[1] Department of Computer Science, University of Manchester
Oxford Road, Manchester, M13 9PL, UK
[2] University of British Columbia, James Hogg iCAPTURE Centre
St. Paul's Hospital, 1081 Burrard St., Vancouver, BC, V6Z 1Y3, Canada
[3] National Center for Genome Resources
2935 Rodeo Park Drive East, Santa Fe, NM 87505, US
[4] Cold Spring Harbor Laboratory
1 Bungtown Road, Cold Spring Harbor, NY 11724, US

Abstract. We have seen an increasing amount of interest in the application of Semantic Web technologies to Web services. The aim is to support automated discovery and composition of the services allowing seamless and transparent interoperability. In this paper we discuss three projects that are applying such technologies to bioinformatics: ^{my}Grid, MOBY-Services and Semantic-MOBY. Through an examination of the differences and similarities between the solutions produced, we highlight some of the practical difficulties in developing Semantic Web services and suggest that the experiences with these projects have implications for the development of Semantic Web services as a whole.

1 Introduction

In the past 10 years, the ability to perform biological *in silico* experiments has increased massively, largely due to the advent of high-throughput technologies that have enabled the industrialisation of data gathering.

There are two principal problems facing biological scientists in their desire to perform experiments with these data. The first of these is distribution–many of the data sets have been generated by individual groups around the world, and they control their data sets in an autonomous fashion. Secondly, biology is a highly heterogeneous field. There are large numbers of data types and of tools operating on these data types. Integration of these tools is difficult but vital. [2]

Biology has coped with this in an effective and yet very *ad hoc* manner. Almost all of the databases and tools of bioinformatics have been made available on the Web; the browser becoming an essential tool of the experimental biologist. The reasons for this choice of technology are partly chance in that the growth in genomic technologies happened to occur contemporaneously with the growth

S.A. McIlraith et al. (Eds.): ISWC 2004, LNCS 3298, pp. 350–364, 2004.
© Springer-Verlag Berlin Heidelberg 2004

of the Web. But many of the key benefits of the Web are also important for biologists. Publishing is economically cheap, technically straightforward, innately distributed, decentralised, and resilient to change. Accessing the Web is likewise simple, requiring no knowledge of specific query languages but enabling "query by navigation" [6].

While this has worked well in the past, it has obvious problems. Many bioinformatics analyses use fragile screen-scraping technologies to access data. Keeping aware of the Web sites on offer is, in itself, a full-time and highly skilled task, mostly because of the complexity of the domain. The application of Semantic Web services to bioinformatics seems a sensible idea as Web services provide a programmatic interface which avoids screen-scraping [14], while semantic descriptions could enable their discovery and composition.

In this paper we describe three architectures, ^{my}Grid, MOBY-Services and Semantic-MOBY, which have been designed to address these problems. All three are aimed mainly at bioinformatics. All three are based on Web or Web-services technologies and use an additional specification of their services to describe the semantics of their operations. All three are high-profile projects in the domain of bioinformatics and come from groups with previous track records of providing solutions for problems of interoperability[1].

The three solutions are also different from each other and from the "idealised" Semantic Web services architecture. In examining these differences, we raise a set of key questions about the applicability of Semantic Web services in practice and present our (partial) solutions for these difficulties.

2 A Day in the Life: Bioinformatics as It Is

Bioinformatics as a discipline has largely grown directly out of the molecular-biology laboratories where it was born. In general, each lab investigated a small region of biology and there are very few labs world-wide working on a single problem. Many of these labs have made their own data available for use on the Web. This data is often un- or semi-structured. Much of the data is composed of DNA or protein sequences, but this has generally been accompanied by large quantities of "annotation"–descriptions (generally in free-text form) of the sources of the sequences, literature citations and the possible function(s) of the molecules. In addition to this raw data, many different tools that operate on it have been developed, most of them with restricted functionality and targeted at performing highly specific tasks.

This situation is slowly changing, largely due to the appearance of large **service providers**, such as the genome-sequencing and -annotating centres. These centres are now increasing their scopes and often provide many different types of information. The primary **service consumer** still remains the small laboratory. Much of the information remains openly accessible.

[1] To our knowledge, at the time of writing these were the only substantial projects using Semantic Web Services within bioinformatics