J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

생물자원 연구데이터의 공동 활용을 위한 데이터 참조모델 개발

Development of a Data Reference Model for Joint Utilization of Biological Resource Research Data

권 순 철[1] 정 승 렬[2*]
Soon--chul Kwon Seung-ryul Jeong

요 약

세계적으로 생물자원 연구데이터는 그 자체로 중요할 뿐만 아니라, 공유되고 활용되어야 한다. 본 논문에서는 명확한 기준 없이 각각의 연구목적과 특성에 따라 개별적으로 구축, 관리되고 있는 생물자원 연구데이터를 공동 활용 할 수 있도록 정보시스템의 구축 단계부터 적용 가능한 데이터 참조모델을 제시한다. 이를 위해 기존 관련 정보시스템의 데이터 모델을 국내외 표준 및 데이터 관리 정책을 기반으로 확장하여 개별 정보시스템에서 공동 활용 할 수 있는 데이터 참조모델을 개발하고 그 적용 절차를 제안한다. 또한, 제안하는 데이터 참조모델의 우수성을 입증하기 위하여 Krogstie의 데이터모델 평가모형을 적용하여 품질수준을 검증하고 국내외 표준들과의 데이터 공유수준을 비교한다. 실험 결과 기존 데이터 모델보다 데이터를 자원, 대상, 활동, 성과의 4단계로 분류하고 엔티티 도출 및 관계를 정의한 데이터 참조모델에서 데이터의 품질과 공유수준이 높게 나타나는 것을 확인 할 수 있었다.

☞ 주제어 : 생물자원, 해양생물자원, 연구데이터, 데이터 참조모델, 정보시스템, 데이터 모델 평가

ABSTRACT

The biological resources research data around the world are not only very critical themselves but should be shared and utilized. Up to now, the biological resources have been compiled and managed individually depending on the purpose and characteristics of the study without any clear standard. So, in this study, the data reference model would be suggested which is applicable in the phase ranging from the start of the construction of the information system and which can be commonly used. For this purpose, the data model of the related information system would be expanded based on the domestic and foreign standards and data control policy so that the data reference model which can be commonly applicable to individual information system would be developed and its application procedure would be suggested. In addition, for the purpose of proving the excellence of the suggested data reference model, the quality level would be verified by applying the Korgstie's data model evaluation model and its level of data sharing with the domestic and foreign standards would be compared. The test results of this model showed that this model is better than the conventional data model in classifying the data into 4 levels of resources, target, activities and performances and that it has higher quality and sharing level of data in the data reference model which defines the derivation and relation of entity.

☞ keyword : Biological Resources, Marine Biological Resources, Research Data, Data Reference Model, Information System, Data Model Evaluation

1. 서 론

생물자원 영역에서 데이터 공유는 국가 간 공동 연구의 필요성과 생물자원의 가치 증대로 인해 매우 중요하게 다뤄지고 있다. 특히 생물자원은 정부·공공 영역뿐만 아니라 학계와 산업계 등의 모든 국가 영역에서 공유의 필요성이 증가하고 있다[1]. 하지만 국내에서 생물자원의 공유는 생물자원에 대한 다양하고 복잡한 연구목적 및 생태적 환경을 기반으로 해야 하는 고유의 특징 및 특성과 생물자원의 데이터에 대한 명확한 기준 및 표준 없이 개별 연구 기관별로 데이터를 정의하고 정보시스템을 구축함으로써 공유목적을 달성하지 못하고 있다.

다양한 목적과 환경을 고려하면서도 정보시스템의 구축과정과 같은 물리적인 데이터베이스 구축과 연계하기 위해서는 보편적인 데이터모델을 공유하는 데이터 참조모델이 매우 유용한 방법이 될 수 있다.

1 Team of Marine Bio-Informatics, National Marine Biodiversity Institute of Korea, Seochun-gun, 33662, Korea
2 Graduate School of Business IT, Kookmin Univ., Seoul, 02727, Korea.
* Corresponding author (srjeong@kookmin.ac.kr)
[Received 17 May 2018, Reviewed 18 May 2018(R2 6 August 2018), Accepted 13 August 2018]

본 연구의 목적은 생물자원 영역의 다양한 연구목적을 가진 기관에서 정보시스템 구축 시 공통적으로 활용할 수 있는 데이터 참조모델을 제시하는 것이다. 제시된 데이터 참조모델은 생물자원 데이터 구축 시 품질수준과 공유수준을 높이는데 기여할 것으로 기대한다.

이를 위해 본 연구에서는 생물자원 영역에서 발생할 수 있는 모든 사건의 기록을 4가지(연구성과, 연구활동, 연구대상, 생물자원) 기준으로 구분하여 상세하게 구체화할 수 있는 방안을 제시할 것이다.

또한, 국내외 생물자원 연구데이터 관련 표준들과 비교실험을 통해 우수성을 검증하고 전문가들에게 설문조사를 통해 적용가능성을 추가 검증할 것이다.

본 논문에서는 생물자원 데이터 영역에서 정보시스템 구축 시 공동 활용할 수 있는 개념수준의 데이터모델인 데이터 참조모델을 제시하고, 각 기관에서 엔티티와 관계를 추가함으로써 논리와 물리 데이터모델로 전개할 수 있는 절차와 방법을 제시할 것이다.

제안되는 데이터 참조모델을 적용한다면 분산된 다양한 특성의 연구데이터들을 빅데이터 등의 최신기술을 활용하여 관련 생태 및 환경 예측과 함께 신종 확보, 해양생물의 보전 및 관리 등의 기초가 될 것이다.

2. 관련 연구

2.1 데이터 참조모델

2.1.1 데이터 참조모델 정의

데이터 참조모델(Data Reference Model)이라고 하면 전사적인 차원의 다양한 요구사항을 정의하고 체계적으로 구조화하기 위한 기틀을 제공하며 내부 정보체계 또는 정보체계 간의 데이터 구조를 표준화하여 향 후 재사용 가능성을 강화하는 것을 가장 큰 목표로 하는 개념적 또는 논리적인 수준에서의 데이터모델이라고 할 수 있다.

더불어 데이터 참조모델은 실제 데이터를 사용하는 사용자 및 이해관계자 등에게 데이터모델에 대한 이해를 강화시킬 수 있으며, 이를 통해 활용 가능한 품질을 보장할 수 있는 데이터모델을 공유하도록 하여 데이터모델의 목표인 재사용성 및 상호운용성 등을 극대화하기 위하여 개발된 것이다[2].

2.1.2 데이터 참조모델 필요성

데이터 참조모델은 어느 특정한 기업체 또는 산업체 등에서 높은 품질의 데이터모델을 공유하고 체계적으로 구축함으로써 실제 정보시스템의 구축 시에 기반이 되는 데이터모델에 대하여 일정수준의 품질을 보장하여 대내외 정보의 공유수준을 강화하기 위한 것이다.

이와 관련된 연구 등에서 제시하는 데이터 참조모델의 필요성 및 제약사항은 다음과 같다.

먼저 Benoit은 데이터 모델러의 기술수준 및 주관적 경험이 아닌 객관적인 데이터 참조모델의 사용이 중요하다는 것을 강조하고 있다.

이는 데이터모델은 업무 및 조직 간의 데이터 공유가 중요한데 대부분 모델러의 주관적 경험 및 사용자 요구사항 등을 기반으로 하는 방법론의 사용 때문에 공유 수준이 낮은 것으로 보았다. 따라서 제한적인 업무영역 내에서 독립적 사용 외에는 모델러에 전적으로 의존하는 방법을 사용해서는 안 된다고 하였다[3]. 즉, 데이터모델은 데이터의 공유 및 공동 활용을 기본으로 장기적으로 활용 범위를 예측할 수 없는 어려운 상황 및 환경에서는 모델러의 주관적 판단을 제한하여야 한다는 것이다.

A. Enders 또한 위 내용과 유사한 주장을 하였는데, 이는 '데이터 모델링이 주관적인 인식으로 계속 발전된다면 실제 특정 모델러의 경험 등에 의해 일정수준의 품질 확보 및 상호간 공유의 어려움이 발생한다는 것이다[4].' 독립적인 상황에서는 발생하지 않거나 고려되지 않았던 품질 문제들이 공유 환경에서 도출되면서 발생함을 지적하고 있다.

앞서 살펴본 관련 연구에서의 결론은 모두가 공유할 수 있는 기준 또는 표준이나 개념적 데이터영역의 객관성을 잃어버리고 특정 모델러의 주관적 견해에 기반하여 모델링을 수행하지 않도록 정보시스템 구축 초기부터 공유 가능한 데이터 참조모델을 사용해야 한다는 것이다.

앞서 살펴본 두 가지의 연구결과들은 본 연구를 시작하면서 연구의 방향성을 결정하는데 매우 큰 근거가 될 것이다. 생물자원 영역의 데이터는 고유한 특성상 공유와 공동 활용이 필수적이기 때문에 개별적 모델러의 경험과 개별기관의 정보시스템 구축목적에 의존하지 않을 수 있는 방법이 모색되어야 하며 개별 모델러의 주관적인 경험 및 판단이 아닌 상호 공유 가능한 데이터모델을 구축하기 위해서 데이터 참조모델이 주로 활용되고 있다.

J. Akoka 등의 연구에서 살펴보면 실제 정보시스템 구축 시 데이터모델링을 함께 수행하는 경우에는 개별 모델러의 기술 수준이 해당 정보시스템의 구축에 아주 큰 영향을 미친다고 하였으며 모델링 시 유사 도메인 또는 업무영역 등의 단위로 데이터 참조모델을 구축·활용한다

면 개발 속도 증가, 개발 위험 감소 등의 효과를 볼 수 있다고 강조하고 있다[5]. 다음은 J. Akoka가 주장한 데이터 참조모델의 장점이다.

첫째, 데이터 또는 공동 활용이 필요한 해당 영역에서 데이터 참조모델은 데이터의 기준 확립 및 표준화와 지속적인 재사용이 가능하게 한다.

둘째, 데이터모델에 대한 기준 확립 및 표준화로 데이터의 이식성, 확장성 등을 강화 할 수 있다.

셋째, 여러 기관(조직) 및 시스템의 상호운용성을 향상시킨다.

넷째, 데이터의 품질 향상과 생산성을 강화시킨다.

다섯째, 데이터모델의 설계 시 중복투자를 방지할 수 있다.

2.2 생물자원 데이터의 특성

2.2.1 공공데이터로서의 특성

국내외에서의 생물자원에 대한 연구 및 결과의 활용은 주로 단기간에 결과를 도출할 수 없으며 오랜 기간 매우 큰 예산이 소요되는 것으로서 대부분 정부차원으로 진행되어지고 있다. 반면 관련 이해관계자는 환경/시민단체, 기업체, 산업체, 연구계, 학계 등 다양하게 구성된다. 즉 생물자원 연구 데이터는 국가 및 공공에서 주로 생산되며 활용자는 산업체, 연구계, 학계 등의 민간이 되는 것이다.

정국환 등의 연구에 의하면 '행정정보 공동 활용, 공공데이터의 민간 개방' 과 같은 환경변화에 따라 이제는 공공의 영역에서만의 데이터가 아닌 민간 영역까지 확대·공유되는 데이터가 되어야 한다고 하였다[6].

또한, 한국정보화진흥원(NIA)에서는 정보시스템의 빠른 변화에 대응하고 데이터의 공유 및 재사용을 강화하여 수요자의 요구를 적극 수용할 수 있도록 핵심적인 역할을 수행하여야 한다고 강조하고 있다[7].

앞서 제시한 바를 종합하면 생물자원 연구데이터가 공공적인 특성을 가지고 있고 데이터의 공개 및 공동 활용을 통해 생물자원 데이터의 가치는 높아질 것이라는 결론을 낼 수 있다.

본 연구에서는 위와 같이 관련 연구결과 및 문헌정보를 통해 제안하는 데이터 참조모델이 데이터를 상호 공유하고 공동 활용하는데 어떤 영향을 미치는지를 그 척도로 활용할 것이다.

2.2.2 생물다양성 기반의 특성

생물자원 데이터는 일반적인 행정, 업무 등의 데이터와는 다르게 '생물의 다양성에 기반하는 데이터'라는 고유한 특징이 있다. 이는 생물자원 데이터에 대하여 매우 넓은 범위의 기준 및 표준화영역을 가지고 있는 TDWG (Taxonomic Databases Working Group)에 의해 정의된 것을 주로 사용하는 특징이 있다[8]. 이러한 특징은 다음과 같다.

첫째, 생물에 대한 개체가 아닌 종(Species)을 기준으로 식별 및 관리하고 있다.

둘째, 종 아래로 종속되는 것으로서 기능, 구조, 허용과 같은 하위 속성들이 존재하고 관리된다.

셋째, 생태 및 환경적 요인 또한 생물자원을 수집·활용하는 그 목적 및 필요성에 따라서 물리적 요소, 생지화학적 요소 등과 같이 다양한 데이터로 구성된다.

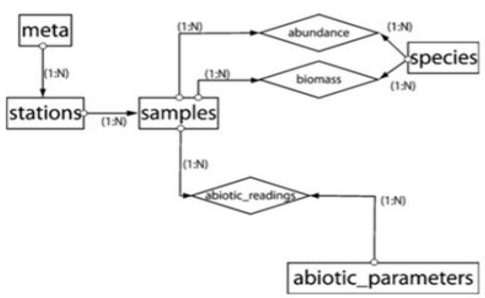
넷째, 산호초, 서식지 등과 같이 생물학의 구조적 구성요소를 위한 특별한 데이터 또는 생태학적 접근 시 표출되는 변화율, 임계 등과 같이 다양한 데이터가 존재한다는 것이다.

다섯째, 타 지역에서 이주되어 서식하는 종을 위한 서식지 관점의 생태 네트워크와 생물의 무결성을 강조하고 있다.

2.2.3 생물자원 데이터의 품질 특성

데이터의 품질적인 측면에서의 생물자원 연구데이터는 종에 대한 분류, 서식 지역, 서식 환경과 같이 세부적인 검증이 가능한 규칙이 추가로 요구된다. Vandepitte는 생물자원영역에서 특별히 요구되는 데이터 품질규칙의 대상과 처리 절차를 세부적으로 제시하였다[9].

그림 1은 Vandepitterk가 주장하는 생물자원 연구데이터의 품질 관리에 대한 절차를 설명한 예시이다. 아래 예시에서 보듯이 생물자원 데이터는 공유할 수 있는 품질규칙이 타 영역에 비해 많고 정교함을 요구한다.



(그림 1) 생물자원 데이터에 대한 품질관리 절차

(Figure 1) Quality control procedures of bio-resource data

결론적으로 말하면 생물자원 영역에서 데이터 참조모델이란 앞서 제시했듯이 생물자원이 가질 수 있는 다양성, 고유성 등의 특성이 고려될 수 있도록 모델링이 수행되어야 한다는 것이다. 즉 객관적인 품질 규칙을 체계적으로 적용할 수 있도록 데이터에 대한 엔티티 및 속성을 도출하여 적용되어야 한다.

2.3 국내외 생물자원 관련 표준

2.3.1 개요

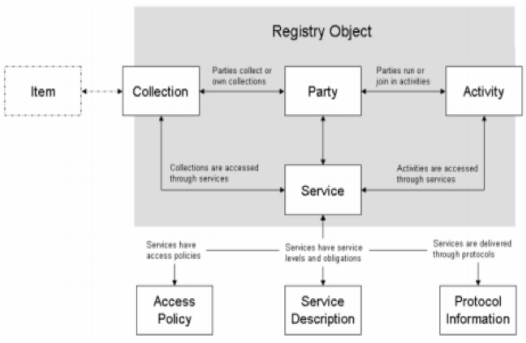
생물자원과 정보서비스를 위한 데이터의 가장 넓은 범위의 구조는 Darwin Core, ISO-2146 등이 있다. Darwin Core에서는 생물자원에 추가적으로 필요한 영상, 음성, 디지털 문헌 등의 정보자원을 데이터로 표현하기 위해 서브 기준으로 Doublin Core를 제시하고 있다[10,11].

이러한 표준들의 조합은 생물자원에 대한 연구 및 관리 등의 활동목적과 다양한 기관의 종류와 상관없이 적용 가능한 메타데이터 표준을 제공한다는 장점이 있다.

그러나 위의 표준들은 선언적 수준 또는 연계 표준의 항목정의 수준이어서 위 표준만을 적용하여 정보시스템 구축 시 다양한 구축형상이 나타나는 단점이 있을 수 있다. 따라서 실제 정보시스템들 간의 공유와 데이터의 공동 활용을 위해서는 해당 정보시스템에 맞도록 데이터 모델 수준으로 구체화해야 한다.

2.3.2 ISO-2146

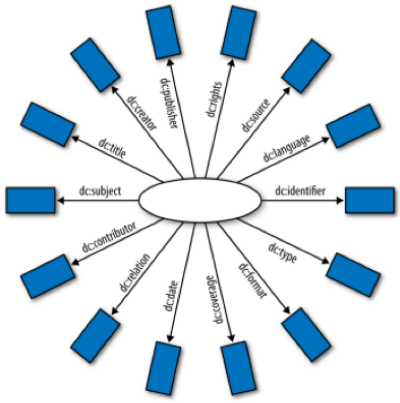
ISO-2146에서는 사용자 관점에서의 정보수집 및 관리와 제공되는 서비스를 일관성 있게 제시하고 있다. 더불어 데이터의 접근정책 등 데이터를 제공함에 있어 필수적으로 필요한 다양한 정책과 그에 대한 명세를 제시하고 있으므로 데이터제공(정보 서비스)에 관련하여 가장 일반적인 표준이라고 할 수 있다. 그림 2는 ISO-2146의 표준 구성 체계를 도식화하여 제공하였다.



(그림 2) ISO-2146 구성 체계
(Figure 2) Configuration system of ISO-2146

2.3.3 Doublin Core, Darwin Core

Doublin Core는 디지털 콘텐츠에 대한 일괄 표준으로 상용목적으로도 폭넓게 사용되고 있는 대표적인 컨텐츠 메타표준이다. Doublin Core의 필수 항목구성은 Title, Creator, Subject 등 총 15개의 요소에 대한 정의 및 설명을 제시하고 있다. 그림 3은 Doublin Core의 구성 체계를 제시하였다.



(그림 3) Doublin Core 구성 체계
(Figure 3) Configuration system of Dublin Core

Darwin Core는 생물다양성 정보과학(Biodiversity Informatics)에 대한 Doublin Core의 확장형으로서 생물다양성 정보에 대한 공유표준으로 Biodiversity Information Standards, TDWG라는 단체를 통해 제공되고 있다. 해당 표준을 사용하는 정보시스템은 GBIF(Global Biodiversity Information Facility)와 OBIS(Ocean Biogeographic Information System) 등이 있다. GBIF, OBIS는 해양생물자원을 포함하는 대부분의 생물자원에 대하여 대표라고 할 수 있는 국제적인 표준이다[12,13].

이에 비해 Darwin Core는 생물자원 영역의 데이터(정보)에 대하여 아주 구체적으로 명확하게 관련 표준을 규정하고 있다. 하지만 데이터를 참조할 수 있는 데이터 참조모델은 제공되고 있지 않기 때문에 Darwin Core만 적용하기에는 한계가 존재한다.

2.3.4 국내 생물자원 정보연계 표준

생물자원 연구데이터와 관련 있는 국내의 표준을 살펴보면 정부의 각 부처별로 생물자원 연구데이터를 연계할 때 필요한 데이터 항목의 형태로 제공된다. 이는 국제적인 표준들을 고려할 때 Darwin Core와 유사한 데이터 구조 및 데이터 항목의 구성을 가지는 것으로서 국내적 표준의 위상 또한 동일한 위치를 가진다고 할 것이다. 다음에서는 환경부와 과학기술정보통신부의 정보연계 표준을 제시한다.

먼저 환경부의 생물자원 연구데이터에 대한 정보연계 표준(항목 구성)은 총 6개로 구분된 정보들의 그룹으로 구성되며 각각의 그룹에 포함된 세부적인 항목(정보)은 표 1과 같다.

(표 1) 환경부 생물자원 정보연계 표준

(Table 1) Standard for bio-resource information collaboration of Ministry of Environment

그룹	포함 정보
연계표준	공통항목, 사진, 그림, 동영상, 표본, 채집지, GPS, 유용성, 유전정보
종목록	분류군, 국명, 문헌정보, 법정보호종, HS코드
종정보	공통항목, 사진, 그림, 동영상
생물자원	공통항목, 표본, 채집지, GPS
유용성정보	공통항목, 분류, 효소
유전정보	공통항목, 핵산서열정보

다음으로 과학기술정보통신부의 생물자원 데이터에 대한 정보연계 항목구성은 총 18개 정보그룹으로 구성되어 있으며 각 그룹에서 포함하고 있는 세부정보는 표 2와 같다.

(표 2) 과학기술정보통신부 생물자원 정보연계 표준

(Table 2) Standard for bio-resource information collaboration of Ministry of Science and ICT

그룹	포함 정보
공통	명칭, 분류, 관리상태
관찰	분포, 관찰상세, 논문, 특허, 부가정보
표본	채집, 분포, 표본상세, 원산지, 논문, 특허, 부가정보
개체	개체상세, 원산지, 논문, 특허, 부가정보
기관	기관상세, 논문, 특허, 부가정보
조직	조직상세, 논문, 특허, 부가정보
배아	배아상세, 논문, 특허, 부가정보
종자	종자상세, 원산지, 논문, 특허, 부가정보
세포(주)	세포상세, 배아, 논문, 특허, 부가정보
균주	균주상세, 논문, 특허, 부가정보
체액	체액상세, 분리원, 논문, 특허, 부가정보
유전자	유전자상세, 논문, 특허, 부가정보
추출물	추출물상세, 분리원, 논문, 특허, 부가정보
핵산서열	핵산서열상세, 분리원, 논문, 특허, 부가정보
발현	발현상세, 논문, 특허, 부가정보
단백질서열	단백질상세, 분리원, 논문, 특허, 부가정보
구조정보	구조상세, 분리원, 논문, 특허, 부가정보
기타	상세, 분리원, 논문, 특허, 부가정보

이상과 같이 환경부와 과학기술정보통신부의 생물자원 연계표준이 다른 것을 확인 할 수 있다. 이는 앞서 제시된 것과 같이 담당 기관(부처)이 대외공유 목적 보다는 개별적인 목적을 가지고 데이터모델을 설계하고 이를 기반으로 정보시스템의 구축이 이루어졌다는 것을 알 수 있다.

2.4 생물자원 데이터 참조모델의 개념적 요구사항

2.4.1 요구사항 정의

앞서 살펴본 관련 연구들의 내용을 종합해 보면 생물자원 영역에서 활용할 데이터 참조모델의 요구사항을 도출할 수 있으며 도출된 참조모델의 개념적 요구사항을 살펴보면 아래와 같다.

첫째, 국가적, 기관(부처)간, 개별적 정보시스템, 다양한 연구계 및 산업체와 연구데이터가 공유·공동 활용되어야 하는 생물자원 연구데이터는 현재 기관별로 수행하는 특수한 고유 업무만을 지원할 수 있는 데이터와 다르게 정보시스템의 초기 구축 시 민간을 포함하는 대국민 대상의 정보서비스를 목적으로 하는 보다 높은 수준의 기준 및 표준화, 공유수준 강화를 전제로 구축되어야 하는 것이다.

둘째, 각각의 생물자원에 대한 개별적인 특정 목적달성을 위하여 구축되어지는 정보시스템들은 실제 구축과정에서 공유를 목적으로 하는 요구사항의 누락 및 생략이 발생할 수 있으므로 국제적 및 국내의 표준 등 전사적 공유의 목적을 달성할 수 있도록 방법과 절차의 제시 및 활용이 필요하다.

셋째, 개별적인 특정업무 및 연구의 목적과는 맞지 않더라도 생물자원 영역의 데이터 공유목적 달성을 위해 반드시 필요한 필수적인 엔티티 및 관계정보는 정보시스템 구축 이전에 제시되어야 한다.

넷째, 제공되는 생물자원 데이터 참조모델을 개별적인 기관에서 실제 정보시스템 구축 시 적용할 수 있도록 체계적인 방법, 절차 등을 제시하여야 한다.

3. 모델 설계

3.1 모델링 기준

3.1.1 데이터 분류 기준

생물자원 영역에서 활용 가능한 데이터 참조모델 제시를 위해 본 연구에서는 생물자원에 대한 연구 및 개발(Research and Development)시 발생되어지는 사건(Event)들을 해석하는 일반적이며 보편적인 기준으로서 4가지 항목(자원, 대상, 활동, 성과)으로 제시한다.

일반적이며 보편적인 기준이라 함은 '모든 생물자원의 연구와 연관되는 사건 및 현상들을 제시된 4가지 기준만으로 설명 가능하다.'는 것이다. 현실세계에서 발생되는 다양한 사건과 현상들을 추상화시킴으로써 단순화하면 최상위로 도출되는 항목들이다. 앞서 제시한 4가지 항목을 구체적으로 정의하면 다음과 같다.

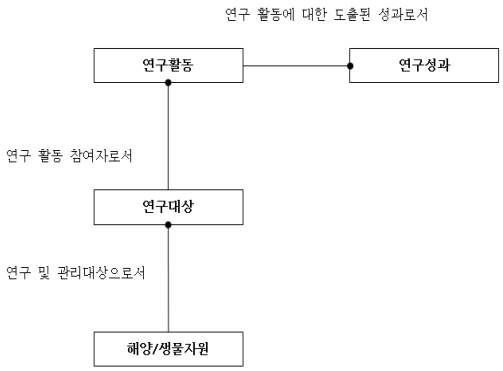
첫째, 자원(생물자원)은 개별 기관 또는 개별 정보시스템의 업무 및 정보시스템 구축목적과 상관없이 관찰의 대상이 되는 생물들로서 계층1로 구성한다.

둘째, 대상(연구대상)은 개별기관 또는 개별 정보시스템의 업무 및 정보시스템 구축목적을 기반으로 하는 주관적인 대상이며, 기관 및 정보시스템의 사용자 그룹에서 일반적으로 사용하는 명칭으로서 계층2로 구성한다.

셋째, 활동(연구 활동)은 생물자원을 가지고 연구·생산·관리하는 일련의 모든 사건기록을 의미하는 데이터이며, 발생 시기(시각, 기간 등)와 같이 객관적 정보만으로 기록하며 계층3으로 구성한다.

넷째, 성과(연구 성과)는 연구 활동에 대한 최종적인 결과로서 파생정보, 성과정보 등을 기록한 데이터를 말하며 계층4로 구성한다.

이처럼 각각의 계층구조와 계층 간의 연관관계를 도식화한 것이 그림 4이다.



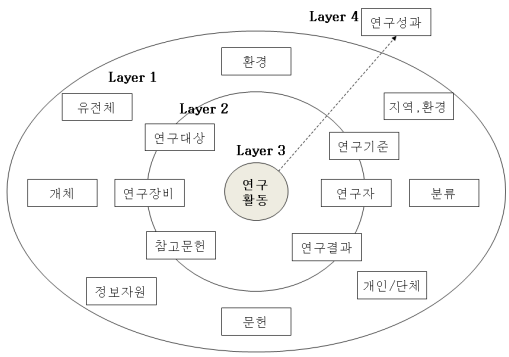
(그림 4) 제안 모델의 개념 구조

(Figure 4) Conceptual structure of proposed model

3.1.2 엔티티 도출 및 관계 정의

엔티티 중 핵심 엔티티의 상호운용성을 극대화하기 위하여 국제적인 표준인 Darwin Core를 기반으로 생물자원 데이터에 대한 엔티티를 도출하였다. 그림 5와 그림 6은 앞서 제시한 계층 구조를 확장하여 생물자원 영역의 핵심 엔티티와 상호관계를 도출하였다.

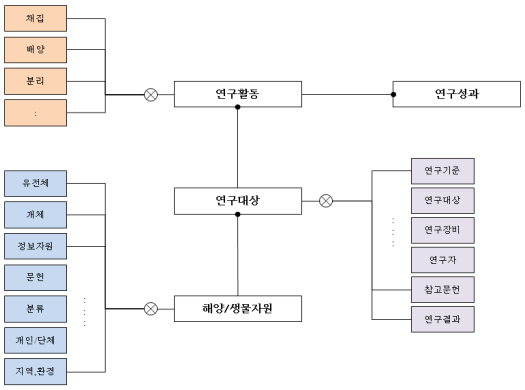
먼저 그림 5에서처럼 실제 도출된 생물자원 데이터의 핵심 엔티티는 Darwin Core의 Class Level의 항목을 기반으로 도출(Class Level Item : 하위에 Item을 포함하는 추상적 명칭으로 엔티티와 속성의 관계와 동일)하였다. 그 후 도출된 엔티티를 각각의 계층에 할당하여 제시하였다.



(그림 5) 핵심 엔티티와 계층 분류

(Figure 5) Important entities and hierarchical classification

그림 6은 도출된 엔티티와 4가지의 계층구조 간 관계를 도식화 한 것이다. 다시 말하면 각 계층수준을 나타내는 엔티티(최초 도출된 4개 기준 엔티티)와 하위의 핵심 엔티티(국제 표준을 통해 도출된 엔티티)는 '상·하위(Sub-Type)'의 '배타적 관계(Exclusive Relationship)'를 가지도록 하였다. 이는 각각의 계층수준에서의 엔티티는 그 하위의 모든 데이터에 대한 식별자역할을 수행하기 때문이다.



(그림 6) 핵심 엔티티와 관계 구조

(Figure 6) Important entities and structure of relationships

3.2 제안 모델링 적용 절차

3.2.1 적용절차 정의

각 정보시스템을 구축할 때 제안된 모델링 방법을 적용하는 절차로는 앞서 도출된 핵심 엔티티의 엔티티-관계도(ERD, Entity Relationship Diagram)를 기준으로 하며 세부적인 절차는 아래와 같다.

첫째, 데이터모델링을 위해 수집된 요구사항에서 엔티티 후보를 추출한다. 엔티티 후보 추출방법은 현재 사용중인 데이터모델링 방법론을 적용하도록 한다. 이는 각각 다른 엔티티 추출 방법론이 활용되더라도 본 연구의 결과에 미치는 영향은 미미할 것이기 때문이다.

둘째, 위에서 제안한 계층구조에 도출된 엔티티 후보를 할당하는 것이다. 또한, 도출된 엔티티 후보를 각각의 계층구조에 할당 시 위에서 제시한 '주관성, 객관성' 기준에 따르도록 한다.

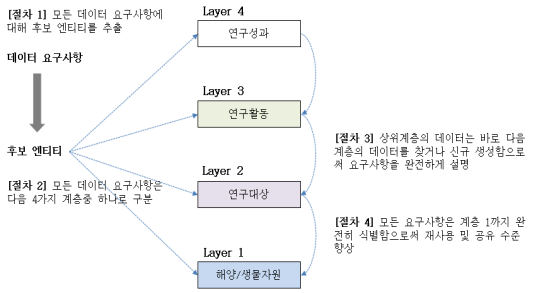
셋째, 엔티티 후보가 각각의 계층(x layer)에 할당되면 각 하위계층(x-1 layer)에 엔티티가 존재하는지를 확인해야 한다. 만약 하위계층에 엔티티가 없다면 추가적으로 새로운 엔티티 후보를 생성해야 한다.(예 : 엔티티 후보 중 '표본'의 경우 '계층2'로 할당하며 '생물개체'의 근거는 '계층1'에서 '생물개체'란 엔티티 후보가 존재하지 않을 경우에 추가하도록 한다.

넷째, 모든 요구사항은 엔티티 후보군의 식별 후 그 상위계층과 연결이 존재하지 않는 하위계층을 찾아 존재하지 않을 시에는 상위계층에 엔티티 후보의 추가 또는 하위계층의 엔티티 후보를 삭제하여야 한다. 이 과정은 모호하게 도출된 요구사항이나 실제 생물자원에 대한 연구활동이 존재하지 않는 경우에 도출되는 엔티티 후보이다. 위에서 제시한 사항은 이를 객관적으로 명확하게 규정하여야 하며 그렇지 않다면 데이터구조에서 제외하여야 한다.

3.2.2 도출절차에 따른 적용 예시

위와 같은 모델링 적용절차는 데이터에 대한 공유와 공동 활용 및 생물자원에 대한 인식수준을 최소한으로 통일시키기 위한 필수적인 요소이다. 이는 실제 정보시스템 구축 목적과 공동 활용을 위한 근거인 생물자원에 대하여 사건기록과 지니는 의미를 서로 분리함으로써 데이터 참조모델에 대한 해석을 명확하게 할 수 있다.

그림 7은 위에서 설명한 데이터모델링 절차를 나타낸 것이다.



(그림 7) 모델링 절차

(Figure 7) Modeling procedure

이러한 모델링의 장점은 위에서 제안한 모델링절차를 '표본(생물표본)' 이라는 실제 데이터에 적용하여 보면 명확하다. 생물자원 데이터 중 표본에 해당하는 데이터는 생물자원에 대한 연구·관리·활용의 핵심으로서 4가지 계층의 활동계층에서도 공통적으로 활용할 수 있다.

다음의 적용사례는 국내 해양생물자원에 대한 통합정보서비스를 제공하는 M-정보시스템에서 추출하였다. M-정보시스템에서 추출된 '표본'이라는 엔티티는 데이터모델을 기반으로 하며 실제 정보시스템의 구축목적을 중심으로 작성되어진 데이터의 구조로 나타난다.

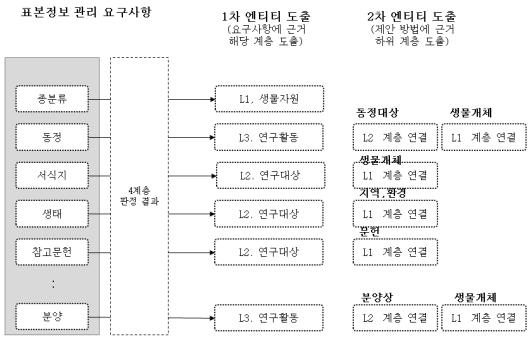
이러한 표본 엔티티 내에는 표본의 종정보, 분류정보, 동정이력, 서식지 정보, 생태적 정보 등과 같이 다양하고 복잡한 형태의 정보를 표본 엔티티 내부에 모두 포함하고 있다. 이렇게 하나의 엔티티에 집적된 데이터는 결과적으로 데이터의 공유 및 공동 활용을 저해하는 장애로

발생될 수 있다.

또한, 정보시스템에 어떤 데이터가 있지를 명확하게 찾을 수 없으며 이로 인해 요청된 데이터가 자신의 DB에서 명확하게 어떤 데이터를 말하는지를 확정할 수 없다.

따라서 본 연구에서는 '데이터를 나누는 기준(표준)에 실제 정보시스템의 구축목적이 담겨 있는지', '생물자원에 대한 사건의 객관적 기록인지 아니면 파장인지'를 기준으로 하고 있다.

그림 8은 '생물자원 중 표본 데이터'를 공유·활용하기 위하여 각각의 엔티티로 식별하는 방법을 나타내고 있다.



(그림 8) 표본 데이터에 대한 엔티티 확장 예시

(Figure 8) Example of entity extension of sample data

4. 실 험

4.1 제안 모델 평가실험 기준

4.1.1 품질수준에 대한 평가실험 기준

본 연구에서 제시하는 데이터 참조모델의 품질수준 측정을 위하여 Krogstie이 제시하는 데이터모델 평가모형을 준용하였다. Krogstie이 제시하는 평가모형은 데이터모델의 품질을 객관적·주관적으로 측정하기 위하여 다음의 8가지 기준을 제시하고 있다[14].

첫째, 정확성(Correctness)은 모델이 데이터모델링에 관한 각종 규칙을 준수하고 있는지를 측정하는 기준이다.

둘째, 완전성(Completeness)은 모델이 시스템의 기능을 지원하는데 필요한 모든 정보를 포함하고 있는지를 측정하는 기준이다.

셋째, 무결성(Integrity)은 데이터모델이 실제 데이터에 적용되는 모든 업무(비즈니스) 규칙을 정의하고 있는지를 측정하는 기준이다.

넷째, 유연성(Flexibility)은 업무(비즈니스) 환경 또는 법규제 변화에 대한 대처가 유연한가라는 데이터모델의 용이성을 측정하는 기준이다.

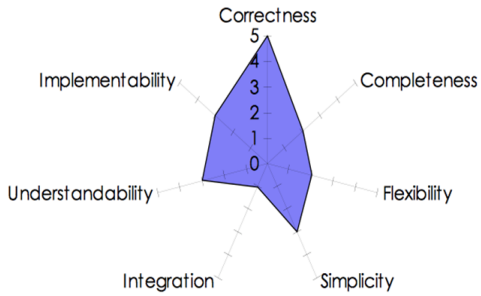
다섯째, 이해가능성(Understandability)은 데이터모델의 개념적인 구조를 모델러가 쉽게 이해하고 적용할 수 있는지를 측정하는 기준이다.

여섯째, 간결성(Simplicity)은 데이터 모델이 도메인 정보를 나타내는 최소한의 엔티티와 관계로 구성되어 있는지를 측정하는 기준이다.

일곱째, 통합(Integration)은 데이터모델이 조직 내외부의 데이터에 대하여 동일하게 적용되는 일관성을 가지고 있는지를 측정하는 기준이다.

여덟째, 구현가능성(Implementability)은 프로젝트 내에서 주어진 시간, 예산, 기술 제약 등에서 데이터모델을 적용한 정보시스템을 쉽게 구축할 수 있는지에 대한 기준이다.

마지막으로 Krogsti의 평가모형에서 중요시 하는 것은 평가기준과 함께 '실제 정보시스템의 구축목적 및 기술능력에 의해 평가기준은 각기 다른 중요도를 가진다.'는 점이다.



(그림 9) OLTP 시스템에서 평가 기준의 중요도

(Figure 9) Importance of evaluation criteria in OLTP system

그림 9는 Krogstie의 연구에서 제시하고 있는 OLTP (Online Transaction Processing) 시스템의 7가지 평가 기준에 따른 중요도를 나타낸 것이다.

본 연구에서는 제안모델이 생물자원 영역의 데이터 참조모델로 적합한지를 측정하기 위해 생물자원 데이터의 특징을 고려하여 평가기준을 선정하고 평가에 활용한다. 생물자원 연구데이터의 특징은 앞서 언급한 것처럼 크게는 '공공성격의 데이터로서 공유의 중요성'과 '생물자원

연구 관점에서의 지원, 관리, 활용'으로 말할 수 있다.

위에서 언급한 두 가지 특징은 모두 고려되어야 하는데 그 이유로는 생물자원의 연구 및 관리를 위한 활동은 정부, 민간, 산·학·연 등의 참여 및 협력을 통한 공동 활용이 필수이기 때문이다. 현재처럼 데이터 공유와 개방 환경에서 생물자원 연구데이터의 수요자(활용자)가 곧 생물자원 연구데이터를 생산하는 주체(OLTP 사용자)가 될 수 있기 때문이다.

위의 내용을 기반으로 업무적인 특성을 가지는 OLTP 시스템과 공동 활용의 특성을 가지는 공공데이터를 모두 가지고 있는 생물자원 연구데이터의 특성들을 고려하여 Krogstie이 제시한 평가기준의 중요도를 재 산정하였다. 표 3은 앞서 제시한 생물자원 연구데이터의 특성을 고려한 평가기준과 OLTP 시스템의 평가기준의 중요도를 비교한 것이다.

(표 3) 평가기준 중요도 비교

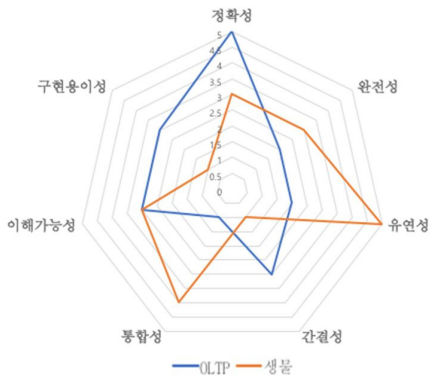
(Table 3) Comparison of importance of evaluation criteria

기준 항목	OLTP	생물자원
정확성	5	3
완전성	2	3
유연성	2	5
간결성	3	1
통합성	2	4
이해가능성	3	3
구현용이성	3	1

표 3에서 보면 OLTP 시스템은 업무처리를 위한 정확성과 정보시스템의 구현용이성이 강조되는 것을 알 수 있으며 공공 데이터의 특성을 가지고 있는 생물자원 연구데이터는 유연성과 통합성이 상대적으로 강조된다. 그림 10은 이러한 두 가지 데이터모델의 평가기준의 중요도 차이를 나타낸 것이다.

데이터모델링에서 공공의 특성을 가지는 데이터와 같이 산·학·연 등 다양한 영역을 지원해야 하는 데이터모델은 종적인 데이터모델링 기법을 사용하고 있는데 이런 경우 일반적인 업무데이터 모델링보다 좀 더 복잡한 Application의 개발노력이 요구된다. 따라서 앞서 제시한 구현용이성, 간결성의 중요도 평가기준은 외부적인 환경의 대응력 강화를 위해 부정평가를 나타낼 수밖에 없다.

또한 국내에서 구축된 생물자원 관련 정보시스템의 구축환경은 제조업, 서비스업 등 타 산업군에 비해 규모가



(그림 10) 데이터 중요도 차이 비교

(Figure 10) Compare data importance differences

작은 기업이 구축하는 사례가 많다. 이로 인해 객관적인 평가를 수행할 수 없으므로 본 연구에서 적용하고자 하는 데이터모델의 품질평가 기준에서는 그 중요도를 낮게 책정할 수밖에 없었으며 일부는 인용할 수 없었다.

이를 제외한 평가기준의 중요도를 높게 산정한 것은 생물자원관리를 위한 업무수행과 공공데이터의 특성을 동시에 만족해야하기 때문이다. 세부적으로 공공데이터라는 것은 관련된 여러 기관 및 정보시스템에서 각각 데이터를 수집하고 활용하려하기 때문에 통합성이 높고, 수집된 데이터를 활용할 다양한 계층을 지원해야하기 때문에 필연적으로 유연성이 강조된다.

따라서 모델러의 주관 및 복잡한 접근이 필요한 구현용이성, 간결성의 기준을 제외하고 정확성, 완전성, 유연성, 통합성, 이해가능성의 나머지 5가지 평가기준을 본 연구에서 제안하는 참조모델의 품질평가에 활용할 수 있도록 선정하였다.

4.1.2 공유수준에 대한 평가실험 기준

제안 모델의 공유수준에 대한 평가는 제안모델이 국가기관 및 국제적인 표준을 제시하는 기관 등에서 요구하는 연계에 필요한 각종 표준과의 부합되는 항목의 충족수준을 평가기준으로 제시하였다. 이는 본 연구에서 제안하는 데이터 참조모델이 기존 데이터모델과 비교하여 연계항목을 더 많이 포함하고 있는지를 측정하는 것이다.

이러한 측정방법을 통해 개별적인 데이터 요구사항을 대상으로 모델링을 수행하며 구축한 데이터모델과 도메인 수준의 대상으로 정보시스템 구축 이전에 구축된 데이터 참조모델간의 데이터 공유수준을 살펴보는 것이다.

1차적으로 본 연구에서 제안하는 데이터 참조모델의 공유 수준 평가를 위해 국내의 생물자원 관련 기관인 환경부와 과학기술정보통신부의 데이터연계 표준의 충족 비율을 측정한다.

4.1.3 평가기준별 실험방법 및 내용

먼저 제안한 데이터 참조모델의 평가는 실험평가와 설문조사 평가의 두 가지를 실시함으로써 실험평가를 통해 객관적인 진단을 하고 설문조사 평가를 통해 주관적인 진단을 병행할 것이다. 이렇게 본 연구에서는 객관적인 평가와 주관적인 평가의 두 가지 평가방법을 모두 실시하여 최종 결론으로 제안하는 데이터 참조모델의 품질수준을 종합적으로 판단할 것이다.

본 연구에서는 데이터모델의 수준을 평가하는 기반 평가 모형으로 Krogstie 모형을 활용하고 있다. 이에 따라 각각의 평가 기준별 세부적인 평가방법 역시 Krogstie 모형을 근거로 실시할 것이다. 다만 설문조사평가 또는 실험평가 시 환경 등의 제약사항으로 세부적인 평가방법의 일부를 조정(삭제, 추가) 하였다.

이를 고려한 평가 기준별 세부 평가방법은 다음과 같다.

평가실험은 정확성, 완전성, 통합성에 대한 평가를 실시하고 유연성과 이해가능성에 대한 평가는 설문조사를 바탕으로 실시한다.

4.2 평가실험 및 설문조사 설계

4.1.1 평가실험 설계

평가실험은 제안된 데이터 참조모델을 기반으로 구축된 데이터모델과 기존에 각 기관에서 적용중인 데이터모델의 품질수준 및 공유수준을 정량적으로 비교할 수 있도록 기준에 근거하여 객관적으로 설계를 하였다.

이를 위해 생물자원을 수집·관리하고 실제 데이터를 대외에 제공하고 있는 국내기관의 'M-정보시스템'과 'e-정보시스템'에 적용된 데이터모델을 비교실험 대상으로 하였다. 첫번째로 'M-정보시스템'은 국내에 분산된 생물관련 다양한 정보를 통합 DB로 구축·연계하여 산·학·연 등 민간에 서비스를 실시하고 있는 정보시스템이며 'e-정보시스템'은 기관 내부에서 생물자원관리를 위한 정보시스템으로 활용하기 때문에 위의 두 정보시스템과의 비교평가를 통해 국내 기관 내부 및 외부용의 정보시스템 특성을 모두 반영할 수 있으며 생물자원이라는 동일한 영역을 대상으로 하기 때문에 본 연구의 실험대상으로 적합할 것이다.

또한 제안하는 데이터 참조모델이 기존 데이터모델을 적용한 정보시스템의 데이터 품질과 공유수준에 기여할 수 있는지를 평가·검증하기 위해 현재 운영 중인 정보시스템의 데이터모델과 제안하는 데이터 참조모델을 모두 적용한 개선 데이터 참조모델의 품질수준 및 공유수준도 함께 비교 실험평가로 실시하였다.

앞서 제시한 바와 같이 실험평가는 크게 데이터모델의 품질수준 측정과 서로간의 데이터 항목이 일치하는지에 대한 데이터공유 수준의 측정으로 실시한다.

먼저, 실험평가에 의한 데이터모델의 품질수준 측정은 앞서 제시한 평가기준에서 첫째 정확성, 둘째 완전성, 셋째 통합성을 평가하여 각각 요구하는 품질기준을 만족하는지를 측정한다.

다음으로 실시하는 데이터공유 수준의 측정은 'M-정보시스템' 및 'e-정보시스템'을 통해 추출된 데이터 항목들이 국내의 생물자원 데이터연계 표준이라고 할 수 있는 환경부, 과학기술정보통신부에서 공개하고 있는 연계표준항목들에 대한 지원수준을 측정하여 산출한다. 환경부와 과학기술정보통신부에서 제시하는 정보시스템 및 데이터 모델은 국가 차원에서 대표적인 데이터연계 표준이라고 할 수 있으며 국내의 생물자원을 바라보는 '보존', '개발/활용'의 관점들을 데이터모델로 제시하고 있기 때문에 본 연구의 실험기준으로 적합하다.

첫째, '정확성 평가'로는 데이터모델링 기술의 각종 규칙을 준수하고 있는지를 평가하는 것으로 데이터모델링 규칙을 위반한 수를 측정한다.

데이터모델링에서 대표적인 규칙은 데이터의 정규화 준수 여부이다. 본 연구에서는 정규화 레벨인 1~5차 정규화 중 본 연구에 제시하는 데이터 참조모델(또는 개념적 데이터모델링) 수준에서 적용할 수 있는 2차 정규화를 적용하여 모델링 규칙의 위배수준을 평가하였다. 2차 정규화를 정의하면 '주식별자(PK) 또는 해당 Key 속성에 종속되지 않는 속성을 분리하는 것'으로써 도출된 엔티티를 명확하게 표현하고 부분적 함수의 종속성과 같은 실제 데이터의 품질수준을 저하 시킬 수 있는 요소를 제거하기 위한 것이다.

둘째, '완전성 평가'로는 데이터모델이 데이터에 적용되는 모든 비즈니스 요구사항을 수용하고 있는지를 평가하였다. 현재 운영 중인 M-정보시스템은 매년 요구사항을 추가적으로 정의하기 때문에 아직 미래의 요구사항이 확정되지 않은 상황이기에 현재 요구사항을 모두 만족하는 것으로 가정하였다. 따라서 제안하는 데이터 참조모델

을 기반으로 하는 개선된 데이터모델은 기존 데이터모델의 보유 데이터 항목(속성)이 제안하는 데이터 참조모델에서 모두 수용가능한지에 대한 판정기준을 완전성으로 하였다. 또한 기존에 적용한 데이터모델에서 사용 중인 실제 데이터항목이 제시하는 개선 데이터모델로 이관 또는 적절하게 폐기된 것이라면 완전성이 확보된 것으로 판정하였다.

셋째, '통합성 평가'로는 실제 적용된 데이터모델이 기관(조직) 외부의 데이터와 일관성을 유지하는지에 대한 평가로서 기관 간의 연계표준에 대한 수용 수준(비율)을 근거로 판정하였다. 이 실험평가는 아래 넷째 항목에서 데이터공유 수준을 평가하는 실험과 함께 이뤄진다.

넷째, '데이터공유 수준' 평가로는 각각 다른 기관 또는 정보시스템에서 동일한 데이터 항목이 공통적으로 존재하고 있는가를 평가하는 것이다. 이는 각각 다른 목적을 가진 정보시스템이라도 실제 현실세계의 동일한 사물(객체)을 그 대상으로 한다면 공유의 근거가 각각의 정보시스템의 데이터모델에 존재해야하기 때문이다.

본 연구에서는 '동일한 생물자원에 대해 각각 정보시스템을 구축한 기존의 두 데이터모델이 환경부, 과학기술정보통신부의 기관(정보시스템)간 정보연계를 위한 표준항목에 부합하는지' 또한 제안하는 데이터 참조모델의 적용 전과 후에 각각의 정보시스템의 데이터공유 수준의 변화수준을 측정하여 그 결과를 제시한다.

개별 연구기관이나 정보시스템에서 다루는 정보항목은 기관 간의 연계항목보다 작은 범위, 규모로 이루어져 있는 것이 일반적이지만 정보시스템별 고유의 데이터 항목이나 내부적인 프로세스가 추가되어 있을 수 있으므로 실제 부분집합의 관계를 가지지는 않는다. 이 실험의 주된 목적은 기관이 개별적인 정보시스템 구축 시 기관 간 및 정보시스템간의 연계표준의 공유와 같은 개념적 데이터모델의 기반에서 모델링이 이루어지고 있는지와 제안하는 데이터 참조모델이 기관의 정보시스템 및 데이터모델에 긍정적 기여가 가능한지를 살펴보는 것이다.

따라서 공유수준에 대한 실험평가는 각각 정보시스템의 동일한 요구사항을 전제하여 기존 데이터모델과 제안하는 데이터 참조모델의 연계표준 부합 수준을 측정한다. 동일한 요구사항을 기반으로 실험평가 대상이 되는 두 데이터모델의 산출을 위해 기존에 운영 중인 정보시스템의 데이터모델에서 제안하는 방법에 의해 추가 개선된 데이터모델을 재 추출하는 방법을 적용하였다.

4.1.2 설문조사 설계

설문조사평가는 '유연성'과 '이해가능성'에 대하여 평가를 수행한다. 설문조사평가는 피 평가자가 제시된 7점 척도 (-3, -2, -1, 0, 1, 2, 3)중에 한 가지를 선택함으로써 수행된다.

먼저 유연성 평가는 데이터모델이 업무(비즈니스)나 규제 변화 등 외부의 환경변화에 대한 대응능력을 측정한다. Krogstie 모형에서 제시하는 유연성 측정의 세부적인 평가기준은 총 3개의 항목으로 다음과 같다.

- Number of data model elements which are subject to change
- Probability adjusted cost of change
- Strategic impact of change

Krogstie 모형을 기반으로 설문조사 문항을 설계한 결과 다음과 같이 총 4개 설문항목이 도출되었다.

첫째, 현재 보유중인 데이터모델이 다음의 변화를 요구 받을 때 모델의 변화수준에 대한 설문으로 총 3개의 문항으로 설계되었다.

둘째, 제안된 데이터 참조모델을 수용했을 경우 정보시스템이 다음의 변화를 요구 받을 때 데이터모델의 변화수준에 대한 설문으로 총 3개의 문항으로 설계되었다.

셋째, 제안 데이터 참조모델을 수용했을 때 도입 및 운영비용에 대한 설문으로 총 3개의 문항으로 설계되었다.

넷째, 제안 데이터 참조모델이 정부와 글로벌 정책을 수용할 수 있는지에 관한 설문으로 총 2개의 문항으로 설계되었다.

이해가능성 평가는 데이터의 개념 및 구조를 관련자(모델러)가 쉽게 이해할 수 있는 능력을 측정한다. 이해가능성 측정을 위하여 Krogstie 모형에서 제시하는 세부적인 평가 기준은 다음과 같이 총 5개 항목 구성된다.

- User rating of understandability
- User interpretation errors
- Application developer rating of understandability
- Subject area-entity ratio
- Entity-attribute ratio

Krogstie 모형을 기반으로 설문 조사항목을 설계한 결과 총 2개 설문 항목이 도출되었다.

첫째, 데이터모델에 대한 직관적인 이해도 관련 설문

으로 총 3개의 문항으로 설계되었으며 현재의 데이터모델과 제안 데이터모델에 대하여 각각 별도로 작성되었다.

둘째, 데이터모델의 부정적 요인을 찾기 위한 설문으로 총 4개의 문항으로 설계되었으며 현재의 데이터모델과 제안 데이터모델에 대하여 각각 별도로 작성되었다.

설문조사 대상선정 및 분류를 위하여 이해가능성에 대한 설문조사평가 수행에 앞서 설문을 작성하는 작성자가 해당 업무의 역할을 선택하도록 하였다. 선택 가능한 역할의 종류는 '생물자원 연구수행자', '생물자원 사업관리자', '시스템 개발자', '시스템 운영자'의 4개 역할이다.

4.3 평가실험 및 설문조사 결과 분석

4.3.1 평가실험 결과

앞서 제시한 바와 같이 실험평가는 데이터모델의 품질수준 및 공유수준의 평가로 진행하였다.

먼저 데이터모델의 '정확도 평가'는 해당 엔티티에 식별자의 정체성과 불일치하는 속성이 존재하는지를 평가하였으며 엔티티마다 몇 개의 엔티티를 추가적으로 포함하는지를 산정하였다. 또한 데이터 참조모델을 기반으로 한 개선 데이터모델에서 2차 정규화 준수 여부를 측정하여 기존 데이터모델과 제안하는 개선 데이터모델의 품질수준을 비교평가 하도록 하였다.

M-정보시스템에서 기존 데이터모델과 개선 데이터모델의 정확도 산정 결과는 표 4와 같다.

(표 4) M-정보시스템 정확도 산출 결과
(Table 4) Output result of precision of M-information system

구분	엔티티 수	식별자도출수	평균식별자수
기존	21	60	2.86
개선	60	60	1.00

e-정보시스템에서 기존 데이터모델과 개선 데이터모델의 정확도 산정 결과는 표 5와 같다.

(표 5) e-정보시스템 정확도 산출 결과
(Table 5) Output result of precision of e-information system

구분	엔티티 수	식별자도출수	평균식별자수
기존	44	113	2.65
개선	113	113	1.00

M-정보시스템과 e-정보시스템의 정확도 측정결과 기존 데이터모델에서는 논리적인 모델링 오류가 발견되었지만 제안하는 개선 데이터모델에서는 논리적인 모델링 오류의 발생이 없었기 때문에 데이터모델의 정확도 측면에서 제안하는 데이터 참조모델을 기반으로 한 개선 데이터모델의 우수함을 도출할 수 있었다.

완전성 평가는 기존 데이터모델의 모든 속성이 데이터 참조 모델링의 방법과 절차대로 개선 데이터모델로 전환될 수 있는지를 평가한다. 실험결과 개선 데이터모델은 기존 데이터모델의 속성을 모두 수용함으로써 완전성 기준을 만족하였다.

완전성 평가결과는 표 6과 같다.

(표 6) 완전성 평가결과
(Table 6) Complete evaluation result

구분	기존 요구사항	수용 요구사항	수용 정도
M-정보시스템	291	291	100%
e-정보시스템	480	480	100%

통합성 평가는 실험설계에서 제시한 바와 같이 공유성 평가와 동일하므로 동시에 진행되었다.

데이터 공유수준 평가는 동일한 속성으로 구성된 두 가지 데이터모델의 엔티티가 해당 연계표준 항목에 얼마나 부합하는지를 측정하였다.

첫째, M-정보시스템의 기존 데이터모델과 개선 데이터모델이 엔티티 수준에서 연계표준 항목과 어느 정도 부합하는 지를 측정하였다.

표 7은 M-정보시스템 데이터모델의 공유수준 평가 결과를 데이터모델의 엔티티 수를 기준으로 비율을 제시하였다.

(표 7) M-정보시스템 공유수준 측정 결과
(Table 7) Measurement result of M-information system sharing level

구분	엔티티 수	공유수준			
		환경부(17)		미래부(24)	
		공유 수	공유비율	공유 수	공유비율
기존	291	1/17	5.88%	4/24	16.67%
개선	480	5/17	29.41%	11/24	45.83%

둘째, e-정보시스템의 기존 데이터모델과 개선 데이터모델이 엔티티 수준에서 연계표준 항목과 어느 정도 부합하는지를 측정하였다.

표 8은 'e-정보시스템' 데이터모델의 공유수준 평가 결과를 제시하였다.

(표 8) e-정보시스템 공유수준 측정결과
(Table 8) Measurement result of e-information system sharing level

구분	엔티티 수	공유수준			
		환경부(17)		미래부(24)	
		공유 수	공유비율	공유 수	공유비율
기존	44	3/17	17.65%	5/24	20.83%
개선	126	12/17	70.59%	11/24	45.83%

이상과 같이 'M-정보시스템'과 'e-정보시스템'이 환경부, 과학기술정보통신부의 표준 연계데이터 항목에 얼마나 부합하는지를 측정 한 결과 위의 두 시스템의 데이터모델 모두에서 제안하는 데이터 참조모델을 기반으로 한 개선 데이터모델의 능력이 우수함이 도출되었다.

데이터 참조모델이 공유 수준에 미치는 영향을 명확히 알아보기 위해 두 시스템의 데이터모델 상호간 공유 수준을 추가적으로 실험하였다.

(표 9) 정보시스템 간 상호 공유 비율
(Table 9) Sharing ratio between information systems

구분	공유 엔티티	M-정보시스템	e-정보시스템
기존	4	19%(4/21)	9%(4/44)
개선	17	28%(17/60)	13%(17/126)

표 9에서 보듯이 환경부 및 과학기술정보통신부의 표준 정보연계 항목의 지원 비율처럼 두 정보시스템 상호간의 데이터 공유수준도 기존 데이터모델에 비해 제안하는 개선된 데이터모델의 공유수준이 높은 것이 확인되었다. 따라서 제안하는 개선 데이터모델이 국내에서 범용적적으로 사용이 가능하다는 공유성을 만족한다고 할 것이다.

4.3.2 평가실험 결과 분석

데이터모델의 품질수준 평가 결과에서 우리는 아래와 같은 결론에 도달할 수 있다.

첫째, 제안하는 데이터 참조모델을 기반으로 하는 데이터모델은 정규화 오류 건수가 두 시스템 모두에서 발생되지 않았기 때문에 정확성(Correctness)을 만족한다. 이것은 제안하는 데이터 참조모델이 데이터의 개념 수준에서 모두 정규화되고 이를 절차에 따라 확장시킬 수 있었는데 이는 개별 모델러의 실수나 임의적인 조치가 최소화되기 때문이다.

둘째, 제안하는 데이터 참조모델이 기존 데이터모델의 요구사항을 모두 수용한다는 것이다. 기존 데이터 항목을 모두 제안하는 데이터 참조모델의 원형에서 수용하거나 실제 모델링 절차를 거쳐 100% 반영되었기 때문인 것이다. 이처럼 범용적으로 제시되는 데이터모델은 기관 및 자신의 특정한 업무에 만족되지 못할 것이라는 우려를 제거하는 것이며 동시에 완전성(Completeness) 평가기준을 만족하는 것이다.

셋째, 제안하는 데이터 참조모델은 내부 업무에 특화된 정보시스템보다 공공과 같이 공유, 개방, 민간 활용을 목적으로 하는 정보시스템에 적합하다. 이것은 두 실험조건 가운데 제안하는 데이터 참조모델의 원형을 사용하는 비율을 보면 명확하게 알 수 있다. 상대적으로 공유, 개방, 민간 활용을 목적으로 하는 M-정보시스템이 내부업무목적의 e-정보시스템에 비해 그 원형에서 데이터 요구를 더 많이 제공하는 것으로 나타나기 때문이다.

넷째, 제안하는 데이터 참조모델은 데이터모델의 공유수준평가 결과에서 해당 정보시스템과 해당기관의 경우 모두 공유수준을 높인다는 것으로 나타났다. 특히 제안하는 데이터 참조모델을 기반으로 할 경우 '환경부, 과학기술정보통신부의 표준 정보 연계에 대한 요구사항의 만족비율이 안정적이다.'라는 것은 제안하는 데이터 참조모델이 공유에 긍정적 영향을 준다는 것을 알 수 있다.

결과를 종합해보면 제안하는 데이터 참조모델의 품질수준을 측정한 정확성, 완전성, 통합성 측면에서 모두 기존 데이터모델보다 우수한 것으로 도출되었다. 이것은 제안하는 데이터 참조모델이 국가적으로 데이터의 연계 및 공유를 통한 활용가능성을 높이는데 기여가 된다는 것으로 결론지을 수 있는 것이다.

4.3.3 설문조사 결과

설문조사는 2017년 9월 15일부터 9월 27일까지 13일간 실시되었으며 세부적인 내용은 다음과 같다.

첫째, 설문 배포의 대상은 생물자원 관련 연구사업의

관리, 수행 등의 경험이 있는 기관(공공) 등에 종사하는 자와 관련 프로젝트에 참여한 경험이 있는 민간의 시스템 관리, 개발 경험자로 선정하였다.

둘째, 배포방법과 설문조사 진행은 우선 배포대상에게 제안하는 데이터 참조모델에 대한 자료를 직접 방문 설명과 관련 교육을 통하여 전달하였으며 일부는 e-mail을 통한 자료 전달과 유선통화로 전달하였다. 또한, 효율적인 설문조사평가를 위하여 구글 설문지를 활용 온라인으로 설문결과를 받아 분석하였다.

셋째, 실제 배포현황은 연구사업 관리자 10명, 연구사업 수행자 20명, 시스템 관리자 10명, 시스템 개발자 20명으로 전체 60명에게 설문지를 배포하였다.

이상과 같이 설문조사를 실시한 결과 총 응답자는 42명이었으며 각각의 실제 응답률은 연구사업 관리자 40%, 연구사업 수행자 100%, 시스템 관리자 50%, 시스템 개발자 65%로 집계되었다.

위의 응답자 가운데 3명이 설문지의 각각 1개 항목에서 무응답 하였으나 전체 평균과 편차로 분석을 실시하는 본 연구의 특성상 해당 무응답자의 전체 항목을 무효처리하는 것 보다 무응답 된 항목만을 평가대상에서 제외하는 것이 타당하다고 판단하여 제외하였다.

실제 응답자의 역할별 응답비율은 생물자원 연구의 수행 및 관리자의 비율이 57%로 정보시스템 관리 및 개발자의 비율인 43%에 비해 높게 나타났다.

연구 분야 담당자는 생물자원에 대한 직접 연구와 함께 생물자원 데이터 표준 연구 혹은 정보서비스에 대한 연구과제 수행자를 포함한 집단이다. 정보시스템 담당자는 생물자원 관련 정보시스템을 구축하는 프로젝트에 참여하여 시스템을 직접 관리, 운영하는 순수 IT 관련 인력이다.

연구사업 수행자가 응답자 중 가장 높은 비율을 차지하는 역할로서 48%의 비율로 나타났다. 생물자원은 주로 국가 주도의 연구과제로 수행되는 것이 일반적이기 때문에 이러한 응답자의 비율은 타당한 것으로 판단하였다.

(표 10) 설문조사 항목별 응답 결과 종합
(Table 10) Response result by survey item overall

항목	평균	편차	응답자수						
			-3	-2	-1	0	1	2	3
5	-1.90244	1.813903	29	0	0	7	2	2	1
12	2.02439	0.757885	0	0	0	1	8	21	11
16	-1.0	1.396424	8	8	7	13	4	1	0

표 10과 같이 설문조사 결과를 항목별로 분석한 결과 5, 12, 16번 항목에 대해 각각 1인이 응답을 하지 않았기에 응답자 수, 평균, 편차의 산출시 제외를 하였다. 더불어 무 응답자는 중복되지 않은 3인으로 구성되어 있었다.

4.3.4 설문조사 결과 분석

설문조사 결과에서 다음과 같은 결론에 도달할 수 있었다.

첫째, 제안하는 데이터 참조모델을 기존 데이터모델과 비교했을 때 제안 데이터모델이 업무(비즈니스) 또는 규제변화에 쉽게 대처할 수 있는 유연성을 가지는 것으로 나타났으며 이는 제안 데이터 참조모델 사용 시 대내외적인 환경변화에 대한 대응력을 갖추는데 긍정적 역할을 한다는 것이며 특히 정책적·전략적 환경 변화 요구를 수용할 수 있는 데이터모델이라는 것이다.

둘째, 제안 데이터 참조모델은 기존 데이터모델 보다 도입비용은 추가되지만 장기적 관점에서는 운영비용이 감소하는 것으로 분석되었다. 이는 기존 데이터모델이 내부 및 외부의 환경변화 대응을 위해 매년 운영비용을 계속 지불하기 때문이다. 따라서 제안 데이터 참조모델은 운영비용을 절감할 것으로 분석되었다.

셋째, 제안 데이터 참조모델은 정부 3.0 등 데이터 개방, 공동 활용에 필요하다는 결과가 도출되었다. 이것은 기존 데이터모델과 비교하였을 때 제안 데이터참조 모델이 정부차원의 공공데이터 개방 및 공유 측면에서 수용가능한 모델이라는 것이다.

넷째, 제안 데이터 참조모델의 엔티티, 속성, 개념적 이해력이 기존 데이터모델 보다 높다는 것이다. 이것은 제안 데이터 참조모델 설계 시 기존 데이터모델처럼 개별 정보시스템이나 특정 업무에 특화된 용어의 사용을 가급적 자제하고 개념데이터 수준에서 단순화시킴으로써 공통적으로 사용되는 분류체계를 기준으로 확장한 결과로 분석된다.

이러한 결과를 종합하면 제안 데이터 참조모델은 정부의 정책 및 환경변화에 대응 가능한 유연성, 소요비용 측면의 효율성, 국가 및 공공 데이터 개방 측면의 수용성, 데이터 참조모델 적용을 위한 이해력 등 모두에서 기존 데이터모델보다 우수하다고 판단할 수 있다.

5. 결 론

본 연구에서는 전 세계적으로 주목받고 있는 생물자원

연구데이터의 구축, 연계, 공동 활용 등을 위하여 국제적인 연계표준과 국내의 연계표준 등에 부합되며 실제 정보시스템 구축 시 활용 가능한 생물자원에 대한 데이터 참조모델을 제안하였다. 구체적으로 본 연구에서는 생물자원 관련 국제 표준인 Darwin Core와 Doubiln Core를 기초로 적용하고 생물자원의 연구목적, 연구방법, 생태환경 등을 추가하여 데이터 참조모델을 4단계의 계층(연구 성과, 연구 활동, 연구대상, 생물자원)으로 설계하여 실세계에 적용 가능한 개선된 개념의 데이터 참조모델을 제안하였다. 제안 데이터 참조모델의 유용성 확인을 위해 실제 운영 중인 기관의 생물자원 관련 정보시스템에 적용하여 국내 연계표준을 적용한 환경부, 과학기술정보통신부의 정보시스템과 비교, 평가 실험을 통해 제안하는 데이터 참조모델이 운영 중인 정보시스템들 간의 연계 및 공유를 위한 요구사항을 매우 만족하는 것으로 나타난 것을 확인 할 수 있었다.

또한, 생물자원 관련 사업 관리자, 담당자, 정보시스템 구축 관리자, 개발자 등 4개의 영역을 대상으로 제안 데이터 참조모델의 유연성, 이해 가능성에 대하여 설문조사를 실시하고 분석한 결과 제안하는 데이터 참조모델이 정보시스템 구축의 요구사항을 매우 만족하는 것을 추가적으로 확인 할 수 있었다.

이처럼 본 연구에서는 국내 및 국외 등 전 세계적으로 다양한 생물자원 연구를 위한 데이터의 수집, 관리, 활용을 위해 구축되는 정보시스템들이 기존의 개별적인 정보시스템에서 나타나는 데이터모델 설계 및 구축으로 인한 정보시스템 간 연계 및 공동 활용의 어려움 해결에 제안하는 데이터 참조모델이 상당한 도움이 된다는 것을 실증적으로 증명할 수 있었다.

또한, 기존 데이터모델로는 다양하게 요구되는 생물자원 관련 요구사항을 수용할 수 없으며 제안하는 데이터 참조모델을 적용하여 정보시스템을 구축한다면 빅데이터 분석을 활용하여 국내외 해양생물자원의 확보, 보전, 관리, 활용에 기여할 수 있을 것이다.

본 연구결과를 바탕으로 향후 본 연구에서 다루지 못했던 다양한 기관 및 정보시스템의 생물자원 관련 연구데이터를 추가하여 실험한다면 제안하는 데이터 참조모델을 보다 정교화 할 수 있을 것이다.

아울러 관련 전문 연구자들의 참여를 확대하여 체계화한다면 보다 범용성을 높일 수 있는 데이터 참조모델로서의 확장 및 활용에 도움이 될 것으로 판단되며 이러한 후속 연구의 추진이 가능할 것이다.

참고문헌(Reference)

[1] Korea Information Center for Agriculture, Forestry & Fisheries, “A Study on the Establishment of Effective Management System for Bio-Resources,” 2010. www.nl.go.kr/app/nl/search/common/download.jsp?file_id=FILE-00008149282

[2] Executive Office of the President of the Unite States, “FEA Consolidated Reference Model Document Version 2.3,” 2007. https://www.reginfo.gov/public/jsp/Utilities/FEA_CRM_v23_Final_Oct_2007_Revised.pdf

[3] B. Otjacques, P. Hitzelberger, F. Feltz, “Interoperability of E-Government Information Systems: Issues of Identification and Data Sharing,” Journal of Management Information Systems / Spring 2007, Vol. 23, No. 4, pp. 29-51. 2007. https://doi.org/10.2753/MIS0742-1222230403

[4] A. Enders, H.D. Rombach, “A Handbook of Software and Systems Engineering: Empirical Observations, Laws and Theories,” Addison-Wesley, Reading, MA, USA, 2003. https://doi.org/10.1109/ms.2004.1270773

[5] J. Akoka, L. Berti-Equille, O. Boucelma, et. Al, “A Framework For Quality Evaluation In Data Integration Systems,” 2007. https://doi.org/10.5220/0002378301700175

[6] Jung Gook-hwan, Moon Jeong-wook, Kwon Sung Mi, “The Future of e-Government : The Future of Public Sector in Korea : Information Sharing,” Korea Information Society Development Institute, 2006. https://www.kisdi.re.kr/kisdi/common/premium?file=1%7C10375

[7] Jeon Jong-su, Oh Dal-su, Yoon Mi-young, “Public Information and Service Activation Strategies in the Gov 2.0 Era,” National Information Society Agency, 2010. http://kiss.kstudy.com/thesis/thesis-view.asp?key=2844124

[8] TDWG, “Bio Information Standards, Darwin Core,” 2014. http://rs.tdwg.org/dwc/2014-11-08

[9] Vandepitte Leen, “Data integration for European marine biodiversity research: creating a database on benthos and

plankton to study large-scale patterns and long-term changes," Hydrobiologia, May2010, Vol. 644 Issue 1, pp.1-13, 2010.
https://doi.org/10.1007/s10750-010-0108-z

[10] ISO TC46/SC4/WG7, "ISO 2146. Information and documentation – Directories of libraries and related organizations," 2010.
http://www.collectionscanada.gc.ca/iso/ill/document/ill_directory/isoawi2146wd13.pdf

[11] DCMI, "Dubline Core Meta Data Element Set Version 1.1," 2012.
http://dublincore.org/documents/dces

[12] Global Biodiversity Information Facility(GBIF), Free and Open Access to Biodiversity Data,
https://www.gbif.org

[13] The Ocean Biogeographic Information System(OBIS), http://www.obis.org

[14] John Krogstie, Quality of Conceptual Data Models., "Practice of Enterprise Modeling: 6th IFIP WG 8.1 Working Conference," PoEM 2013, Riga, Latvia, November 6-7, 2013, Proceedings; 2013, pp.39-53, 15p, 2013.
https://core.ac.uk/download/pdf/52112318.pdf

[15] NSF, "Digital Research Data Sharing and Management," 1, 8. September, 2012.
http://www.nsf.gov/nsb/publications/2011/nsb01211.pdf

[16] P. Spyns, Y. Tang, R. Meersman, "An Ontology Engineering Methodology for DOGMA," Applied Ontology, Vol. 3, No. 1-2, 2008.
https://dblp.org/rec/journals/ao/SpynsTM08

[17] E. Marcos, A. Marcos, "A philosophical Approach to the Concept of Data Model: Is a Data Model, in Fact, a Model ?," Information System Frontiers, Vol. 3, No. 2, 2001.
https://doi.org/10.1023/A:1011460711754

[18] R. Valverd, M. Toleman, "Ontological Evaluation of Business Models: Comparing Traditional and Component-Based Paradigms in Information Systems Re-Engineering," Ontologies Integrated Series in Information System, Vol. 14, 2007.
https://doi.org/10.1007/978-0-387-37022-4_3

[19] US Gov Accountability Office, "The Challenge of data sharing : Results of a GAO-Sponsored," Oct 20, 2000.
https://books.google.co.kr/books?isbn=1428970401

◐ 저 자 소 개 ◑

권 순 철(Soon-chul Kwon)
2010년 한국산업기술대학교 컴퓨터공학과 졸업(학사)
2012년 인천대학교 정보기술대학원 졸업(석사)
2018년 국민대학교 비즈니스IT전문대학원 졸업(박사)
2015년~현재 국립해양생물자원관 해양생명정보부
관심분야 : 프로젝트관리, 정보자원 관리, 데이터 분석 및 활용 etc.
E-mail : kwonsc75@mabik.re.kr

정 승 렬(Seung Ryul Jeong)
1985년 서강대학교 경제학과 졸업(학사)
1989년 미국 위스컨신 대학교 경영정보학과 졸업(석사)
1995년 미국 사우스케롤라이나 대학교 경영정보학과 졸업(박사)
1997년~현재 국민대학교 비즈니스IT전문대학원 교수
관심분야 : 정보시스템 구현, 프로세스 관리, 프로젝트 관리, 정보자원 관리 etc.
E-mail : srjeong@kookmin.ac.kr