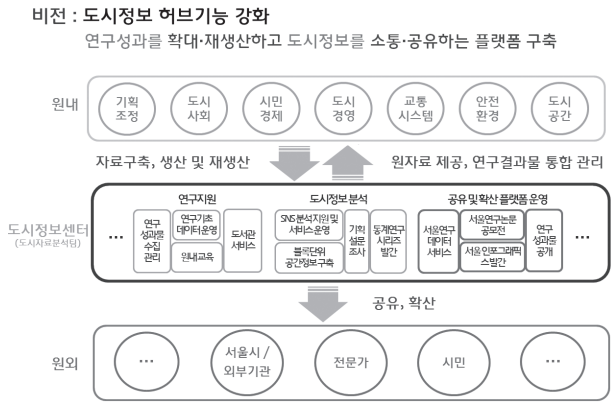
서울연구원 도시데이터 구축·확산의 경험과 과제

이정호, 최선희 _ 서울연구원

1. 서론

서울연구원은 정보의 소비자이자 생산자이다. 다양한 학제간의 연구를 수행하면서 실제 자료에 기초한 현장중심의 연구를 추구하고 있는 만큼 여러 분야의 도시데이터가 필요하며, 또한 연구를 수행하는 동안 다양한 데이터 및 연구수행의 결과물을 자체 생산하기 때문이다. 이와 같이 연구기관은 그들의 연구 활동과정에서 연구데이터를 소비하고 다양한 연구 성과물을 생산하고 있다.

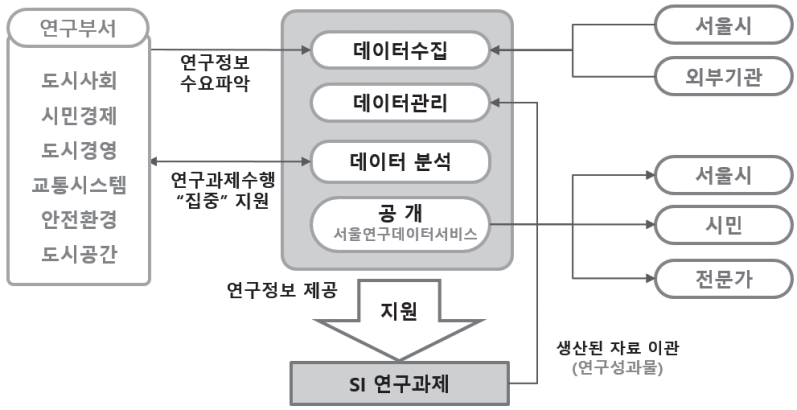
도시정보의 허브기능을 담당하는 도시정보센터의 주된 역할은 연구성과를 확대·재생산하고 도시정보를 소통·공유하는 일이다. 과거에는 도시데이터 수집 및 구축·운영의 목적을 원내 연구과제 지원으로만 한정하였으나, 최근에는 연구원이 축적한 다양한 연구자료 및 도시정책 관련 공공데이터 공개 요구에 대응하여 공공기관, 전문가, 시민, 기업들과 공유할 수 있는 플랫폼을 구축하여 이를 통해 서비스를 제공하고 있다.



〈그림 1〉 도시정보센터 비전과 역할

2. 도시데이터의 구축 · 확산

서울연구원은 서울시를 포함한 여러 기관과의 연계를 통하여 도시계획, 교통, 환경, 산업·경제, 문화, 복지 등 다양한 분야의 행정·통계 데이터를 수집하여 연구에 활용하고 있다. 그러나 수집된 데이터는 다양한 목적과 형태로 연구와는 무관하게 수집되고 있어 도시를 이해하기 위한 자료로 가공되거나 지표화 할 필요가 있으며, 도시데이터의 특성상 통계 데이터와 공간 데이터의 접목이 필수적이다. 이에 따라 도시정보센터에서는, 데이터의 수집·관리·분석절차에 따라 도시데이터를 구축하고 원내 연구과제에 지원하는 업무를 수행하고 있다. 최근에는 서울시, 시민, 전문가 등 대외공개가 가능한 도시정보 공유플랫폼(서울연구데이터서비스)을 별도로 운영함으로써 도시데이터 및 연구성과의 확산을 위한 노력을 기울이고 있다.



〈그림 2〉 데이터 수집·구축 및 지원 체계

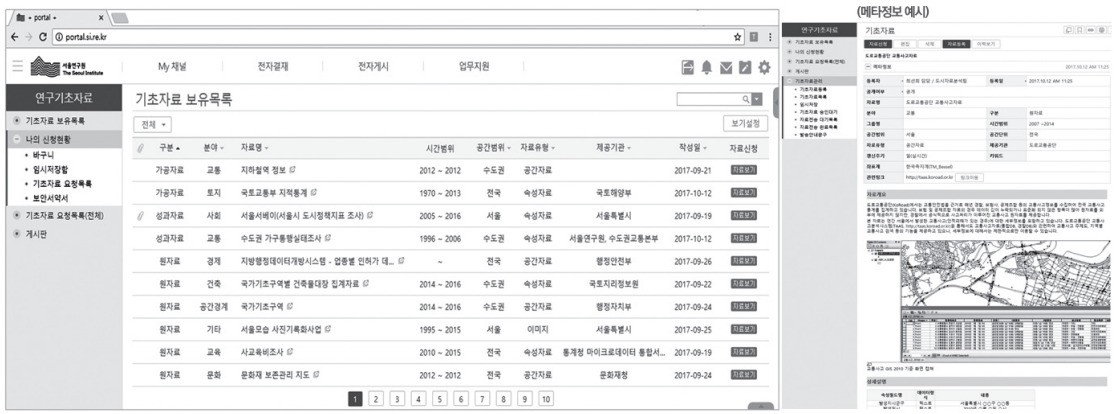
1) 데이터 수집

원내 연구정보 수요조사 결과를 반영한 연간 수집계획에 따라서 서울시 및 외부기관에서 생산하는 데이터를 체계적으로 수집하고 있으며, 이중에서도 특히 연구에 활용도가 높은 데이터는 정례화하여 시계열로 구축하고 있다. 현재 인구, 주택, 토지, 교통, 산업경제, 환경, 도시계획 등 10개 분야 120여종의 데이터를 보유하고 있으며, 실제로 많은 연구과제에서 이를 활용하고 있다.

2) 데이터 관리

수집데이터는 토지관리, 건축물관리, 도로·교통관리 등 다양한 목적의 상이한 시스템에서 제각각의 형태로 생성·구축되어 있기 때문에 데이터 구조를 파악해서 형식을 맞추는 등의 가공작업이 필요하다. 또한 데이터의 오남용을 방지하기 위해서 데이터의 출처, 상세설명, 변동이력, 사용시 유의사항 등 상세한 메타정보를 작성하여 관리할 필요가 있다. 수집데이터의 원본은 별도로 보관·저장하고, 가공작업을 완료한 데이터와 메타정보는 연구기초자료시스템(원내포탈)의 콘텐츠로 구축하여 연구지원을 목적으로 운영·관리하고 있다.

연구원내 연구성과물 중에서도 연구보고서 및 결과데이터, 연구과제 설문조사 자료 등 활용가치가 높은 데이터는 아카이빙을 통해서 체계적으로 관리하고 있다. 연구성과에 대한 외부공개 요구가 많아지면서 성과물의 체계적인 수집·관리·공개 절차가 더욱 중요해지고 있다.



〈그림 3〉 연구기초자료시스템(원내포탈)

3) 데이터 분석

대용량·시계열 데이터는 DBMS로 구축함으로써 분석의 효율성을 높일 수 있다. 또한 데이터 정제작업을 통해 분석결과의 정확도를 높일 수 있고, 통계데이터를 공간데이터와 접목해서 다양한 융합분석을 시도할 수 있다. 이처럼 데이터를 분석하기 쉽게 재구축하거나 연구에 활용할 수 있도록 가공하는 등 전처리 과정

을 계획적으로 수행함으로써 원활히 원내 연구과제를 지원하고 있다.

또한, 별도의 요청이 있는 경우 분석방법 및 분석 툴 사용에 대한 기술지원을 수시로 제공하는 한편, 연구역량 강화를 위해서 오픈소스 프로그램 활용 및 자료 분석 방법에 대한 교육도 병행하고 있다.

데이터 구축형 연구과제의 경우에는 연구진으로 참여하여 "집중지원"을 제공하기도 하고, GIS 공간정보 분석 및 텍스트마이닝, 소셜데이터 분석 등 자체 분석 업무를 수행하고 있다.

4) 데이터 공개 (서울연구데이터서비스, http://data.si.re.kr)

데이터 개방·공유가 활발히 이루어지고 있는 시대적 요구에 대응하여 서울연구원이 축적한 다양한 도시데이터 및 연구성과를 공개할 수 있는 서울연구데이터서비스를 구축하여 3년째 운영하고 있다. 도시데이터를 구축·운영 및 연구에 지원하면서 자연스럽게 축적된 도시정보센터의 역량과 노하우를 바탕으로 공급자 위주의 일방적인 정보제공이 아닌, 사용자가 이해하기 쉽고 사용하기 편한 맞춤형 정보제공을 위해 노력을 기울여 왔다. 데이터가 중심이 되고 데이터 공유가 일상이 된 현재, 제공하는 데이터의 양은 그리 중요하지 않은 듯하다. "어떤 정보를 어떻게 제공할 것인가"에 대한 고민이 무엇보다 더 중요한 시점이다.

3. 서울연구데이터 서비스 운영

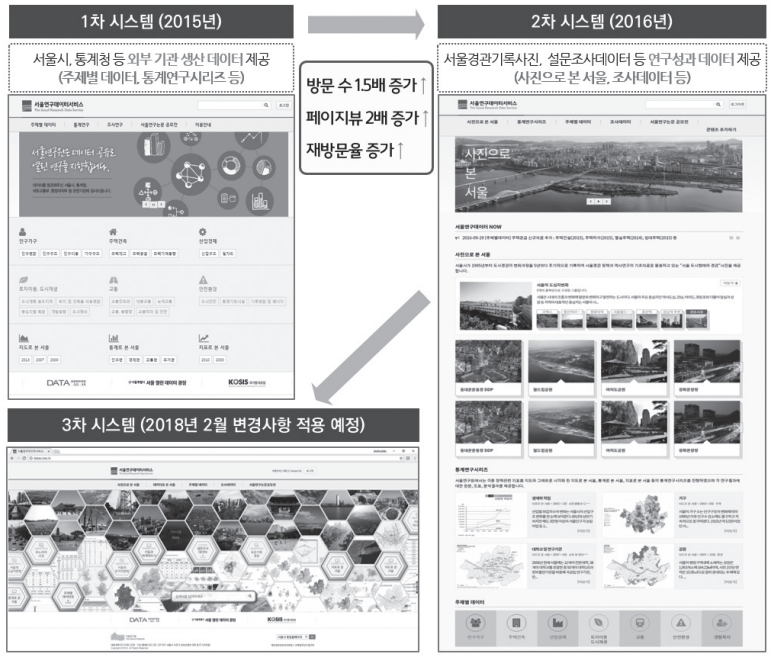
1) 개요

여러 기관들의 공공데이터 개방을 위한 노력으로 양적 성장은 이루었으나, 수요자가 원하는 정보를 제공하지는 못하는 실정이다. 개인정보보호의 이유로 일부 데이터만을 공개하거나, 전국·시도·자치구 등 대단위 지역의 통계데이터 만을 제공하는 경우도 많아서 활용에 많은 제약이 따른다. 또한, 데이터의 설명이 부족하고 정제되지 않은 자료를 제공하는 경우가 있어 데이터의 특성을 제대로 파악하기가 어렵고, 자료의 오남용에 따른 연구결과의 신뢰도 저하가 우려되는 것이 사실이다.

서울연구데이터서비스는 연구 분석에 활용도가 높은 도시데이터를 정책관련

주제별 지표로 분류하여 사용자가 이용하기 편리하도록 가공하여, 상세한 메타정보와 함께 제공하고 있다. 이는 올바른 데이터 활용에 대한 가이드를 제시함으로써 사용자의 이해도를 높이고, 그 결과로 양질의 연구성과를 확보하기 위함이다.

2) 시스템 구축 현황



〈그림 4〉 서울연구데이터서비스 구축 현황

서울연구데이터서비스는 콘텐츠 관리가 용이한 오픈소스 CMS 드루팔(drupal)을 기반으로 시스템을 구축하였으며, 서울시 등 외부기관에서 협조 받은 행정·통계 데이터를 정책관련 지표로 분류하여 가공·정리한 "연구기초데이터"를 주요 콘텐츠로 내세워 2015년 5월 서비스를 처음 개시하였다.

이후, 연구성과 확대·재생산에 대한 중요도의 증가 및 연구원 자체 생산 성과자료의 공개 확대에 대한 원내외 요구가 높아짐에 따라, 신규 콘텐츠의 발굴 및 시스템 개선사업을 추진하여 2016년 7월에 2차 서비스(개선)를 개시하였다. 서울의 과거와 현재 모습을 담은 서울경관 디지털사진 및 연구과제 설문조사데이터 등 연구성과 데이터를 신규 콘텐츠로 제작하여 서비스를 제공하면서, 일일 사용자는 약 250명으로 전년도 대비 1.5배 증가하였고 페이지뷰도 2배 증가하는

등 소정의 성과를 이루었다.

메인페이지는 사용자가 사이트에 방문했을 때의 첫인상이므로 서비스 콘텐츠의 특성을 드러내거나, 사용자의 호기심 내지는 흥미를 이끌어낼 수 있어야 한다. 이러한 점에 착안해서 디자인 및 기능개선을 진행하였고 2018년 2월, 3차 서비스(개선)를 개시할 예정이다.

3) 주요 콘텐츠 및 특성

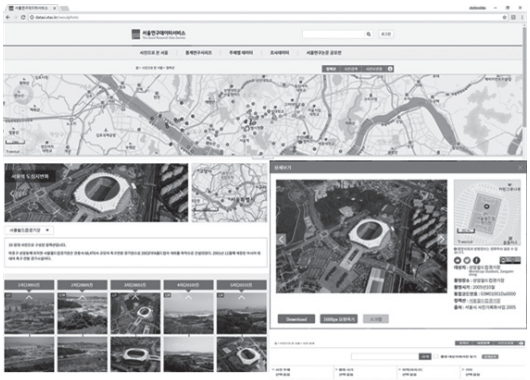
서울연구데이터서비스에서 제공하는 대부분의 콘텐츠는 보안 등과 관련이 적어서 회원가입 없이도 간단히 데이터를 다운로드 할 수 있다. 주요 콘텐츠로는 사진으로 본 서울(1995년 이후 서울시 경관사진), 조사데이터(서울연구원 연구과제의 설문조사 데이터), 통계연구시리즈(2000년 이후 서울연구원에서 발간한 통계·지도 자료집), 주제별 데이터(연구 분야별 통계데이터 모음) 등으로, 데이터 현황은 아래 표와 같다.

〈표 1〉 주요 콘텐츠 현황

주요 콘텐츠	내용
사진으로 본 서울	· 제공자료: 1995년~2015년(5년 단위) 서울경관기록화 사진 25,000건 · 파일유형: 이미지(JPG) · 이용방법: 900/1600픽셀 이미지 검색 및 직접 다운로드
조사데이터	· 제공자료: 서울연구원 연구과제 위탁설문조사 데이터 20여 종 및 서울복지실태조사(2015) 등 서울연구원 조사연구시리즈 3종 · 파일유형: 엑셀(XLSX), 텍스트(TXT), 문서(PDF) · 이용방법: 조사표, 보고서, 마이크로데이터 등 직접 다운로드
통계연구시리즈 (아카이브형 연구과제 데이터)	· 제공자료: 서울연구원에서 발간한 통계·지도 자료집 및 아카이브형 연구과제 데이터(서울과 세계대도시, 지도로 본 서울, 통계로 본 서울, 지표로 본 서울 등) · 파일유형: 원문(PDF), 이미지(JPG, PNG), 엑셀(XLS) · 이용방법: 원문 PDF 및 본문, 이미지, 표 등 웹 콘텐츠, 데이터 직접 다운로드
주제별 데이터	· 제공자료: 서울시, 통계청 등 공공기관에서 공개하는 데이터를 상세한 메타정보와 함께 재정리한 연구 분야별 통계 정보 10종 330건 · 파일유형: 엑셀(XLSX), 텍스트(TXT) · 이용방법: 직접 다운로드

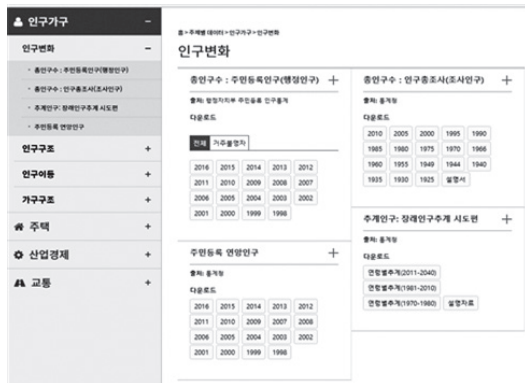
사진으로 본 서울

· 고화질 사진 다운로드, 서울의 변화 컬렉션, 검색기능 제공



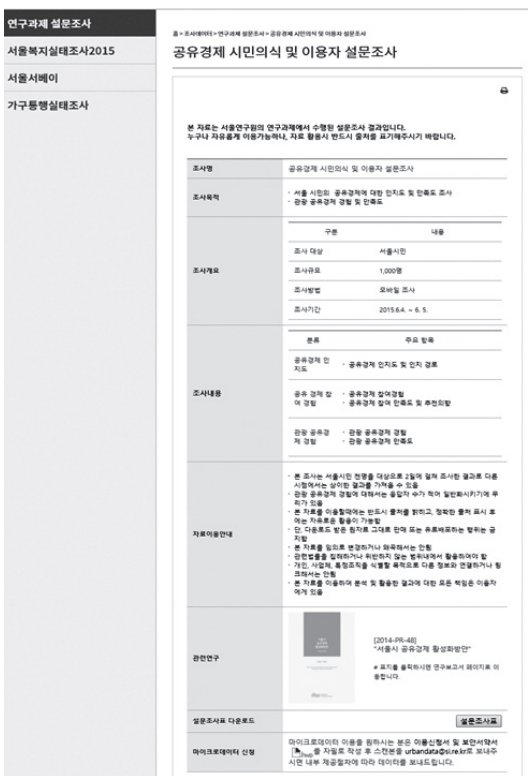
주제별 데이터

· 통계데이터 다운로드 및 상세한 메타정보 제공



조사데이터

· 서울연구과제 설문조사 메타정보, 데이터 다운로드 제공



통계연구시리즈

· 원문 PDF 및 본문, 도표, 이미지, 데이터 다운로드 제공

〈그림 5〉 서울연구데이터서비스 주요 콘텐츠 및 특성

서울연구데이터서비스 콘텐츠의 가장 큰 특징은 수요자 맞춤형 데이터를 제공한다는 점이다. 서울의 현재와 과거를 비교할 수 있는 디지털경관사진을 비롯하여 서울연구원이 수행하였던 “지도로 본 서울”, “지표로 본 서울” 등 연구결과 원문(PDF), 원문에 포함된 그림과 그래프를 고화질 이미지로 내려 받아 활용할 수 있으며 관련 데이터 또한 엑셀(XLSX)형식으로 내려 받아 새로운 분석에 활용할 수 있다. 특히, 전문가를 위해서는 인구·가구, 주택, 산업·경제, 교통 등 다양한 분야의 도시데이터를 읍면동·국가기초구역 등 소지역 단위로 내려 받을 수 있도록 제공하고 있다. 또한 자료의 이용방법 및 이용 시 주의사항 등을 메타정보에 포함하여 제공함으로써 이용자들이 자료를 정확히 이해하고 활용할 수 있도록 안내하고 있다.

향후, 이용자 요청에 즉각적으로 반응하는 인터랙티브 맵, 동영상 등 다양한 시각화 콘텐츠를 추가하고 주기적인 데이터 작성 및 안정적인 시스템 관리를 통해서 콘텐츠 확충 및 서비스 고도화를 지속적으로 추진할 계획이다.

4. 결론 및 제언

최근 각 기관별로 연구데이터 관리가 중요한 이슈로 떠오르고 있다. 공공도서관은 논외로 하더라도, 대학도서관은 교수 등 학내 구성원들의 성과 관리를 위해, 연구기관 전문도서관은 본원이 생산한 데이터의 자산화 및 활용이 필요하기 때문이다. 연구기관의 경우 이미 오래 전부터 연구데이터를 포함한 성과물의 수집 및 효율적 관리, 연구에 재활용하는 방안에 많은 공을 들여왔다. 연구성과에 대한 외부공개·공유에 대한 요구가 증가하면서 연구데이터의 수집·관리가 더욱 중요해진 시점이다. 서울연구원 또한 많은 시행착오를 거치며 현 단계에 이르렀다. 그럼에도 불구하고 아직도 운영상의 문제점과 개선해야 할 과제가 많이 남아있다.

첫째, 연구데이터의 정의 및 운영 목적을 명확히 해야 하고 이에 대한 기관 내 합의가 있어야 한다. 연구기관에서 연구데이터관리가 어려웠던 이유는 지금까지 연구데이터는 연구자 개인의 성과로 간주해서 연구자로부터 제출의무를 강제할 수 없기 때문이다. 또한 연구 과정 중에 생산되는 성과물의 종류와 포맷이 다양하여 수집 대상 자료의 유형과 범위를 한정하는데 한계가 있다. 따라서 연구데이터의 수집 대상 자료의 정의, 수집된 데이터의 관리를 위한 연구원 내부의 체계가 필요하다. 특히 연구데이터 납본에 대한 인식이 과거보다는 눈에 띄게 당연시 되

고 있는 상황이므로 연구데이터 구축 목적이 명확하다면 과거처럼 데이터 수집 단계에서의 어려움은 없을 것이다.

둘째, 연구데이터 관리 표준이 필요하다. 연구데이터는 연구 과정 중에 또는 종료 후 생성된 데이터로서 다양한 포맷으로 존재할 수 있다. RDB, 텍스트파일 등 비정형데이터, 도면 등 이미지파일, GIS 공간데이터 등 다양한 포맷을 담을 수 있는 시스템과 메타데이터 정의 툴이 필요하다. 연구데이터는 비교적 상세한 메타데이터 명세 작성 기능과 버전 관리 기능이 필수적이다. 따라서 기존 LAS를 활용하기에는 한계가 있다.

셋째, 연구데이터와 관련된 기술 습득이 필요하다. 우리는 도서에 대한 관리 기술에 대해서는 전문가이지만 각 분야별 데이터에 관해서는 문외한이다. 따라서 데이터를 이용자에 요구에 맞게 제공하기 위한 데이터 정제기술, 이용자의 문의에 대응하고, 자관 내외에서 활용할 콘텐츠를 만들 수 있는 데이터 분석기술 및 다양한 응용프로그램 운용 능력이 필요하다.

Open Acess를 필두로 Open Science, Open Data, Open Research 등 데이터 관련 용어가 화두가 된 요즈음 연구기관에 근무하는 전문도서관 사서로서 혼란스럽기만 하다. 특히 1인 사서로 운영되는 전문도서관이 많은 국내의 상황에서, 연구데이터에 관한 이슈가 도서관 업무에 부담이 될지 기회가 될지 궁금 하다. 각 기관들은 자관이 생산한 데이터에 관심을 갖게 되었고, 많은 공공 기관들이 데이터 개방에 참여하고 있다. 누군가는 데이터를 생산하여 데이터가 모아지고, 누군가는 관리하고 또 누군가는 그 데이터를 활용할 것이다. 4차 산업혁명은 데이터 혁명이기도 하다. 혹시 이것이 우리가 가야할 새로운 길이 아닐까 생각해본다.